

JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

G. THÉODORE

Préoccupations actuelles en matière d'organisation de la collecte d'informations numériques

Journal de la société statistique de Paris, tome 115 (1974), p. 19-51

http://www.numdam.org/item?id=JSFS_1974__115__19_0

© Société de statistique de Paris, 1974, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

III

ARTICLES

PRÉOCCUPATIONS ACTUELLES EN MATIÈRE D'ORGANISATION DE LA COLLECTE D'INFORMATIONS NUMÉRIQUES

After a summary of the French organisation of the statistical information gathering, we indicate the actual factors which influence the nature of this gathering. We point out the problems to solve and how we try to improve data quality.

Sich auf die Organisation der Sammlung statistischer Informationen in Frankreich beziehend, wurde von den kurzlich ins Leben gerufenen Faktoren gesprochen, welche die Natur dieser Kollekten beeinflussen. Die gestellten Probleme wurden besser umrissen, sowie auch was gemacht wurde um die Qualität der gesammelten Nachrichten zu verbessern.

Despues de un recuerdo sobre la organización del aparato encargado de la colecta de l'informacion estatistica en Francia, se ha indicado cual eran los factores recientes que han influido en la natura de esta colecta. Se han senalado los problemas que se presentan y la obra emprendida para mejora las validades de los datos.

A — RAPPEL SOMMAIRE DE L'ORGANISATION DE L'APPAREIL CHARGÉ DE LA COLLECTE DE L'INFORMATION STATISTIQUE

1. *L'Institut national de la statistique et des études économiques* (I. N. S. E. E.) a été créé par la loi du 27 avril 1946. Il est né de la fusion de l'Institut de conjoncture fondé en 1938 et du Service national des statistiques qui réunissait depuis 1941 la Statistique générale de la France et le Service de démographie ⁽¹⁾. A cette époque, les missions assumées étaient les suivantes :

- Coordination statistique pour l'ensemble de l'Administration.
- Exécution d'enquêtes particulières.
- Analyse des résultats, présentation et publication.
- Centralisation de la documentation statistique élaborée par d'autres unités.
- Gestion de fichiers de personnes et d'unités économiques.
- Études de la situation économique.

1. 25^e anniversaire de l'I. N. S. E. E., supplément au n° 24, *Économie et Statistique*, juin 1971.

Depuis 1962, l'I. N. S. E. E. a été chargé en plus des tâches précitées de :

- L'établissement des comptes nationaux rétrospectifs.
- L'apport de son concours technique à l'élaboration des comptes nationaux prospectifs établis pour la préparation des budgets économiques.

L'I. N. S. E. E. est décentralisé :

Il comprend une Direction générale à Paris (conception, coordination des travaux, exécution des études et des recherches) et diverses unités régionales qui exécutent les travaux prescrits à l'échelon national et les travaux régionaux nécessaires aux études régionales et à la mise en œuvre des programmes d'action régionale ainsi qu'à la diffusion des informations.

Ces unités régionales se composent de dix-huit Directions régionales, trois Services régionaux, quatre Services départementaux dans les départements d'outre-mer, six Centres nationaux informatiques, deux Centres nationaux de traitement.

2. *La déconcentration de l'I. N. S. E. E.* — Depuis 1945, une politique a été suivie d'implanter dans chaque ministère technique, une cellule statistique dont la tâche principale est de rassembler les séries statistiques disponibles ou d'en créer par le biais d'enquêtes ou de réorganisation des circuits administratifs ⁽¹⁾. Cette implantation initiale est actuellement très développée, mais pour certains départements ministériels, elle demeure récente.

L'action générale de coordination de l'I. N. S. E. E. ⁽²⁾ s'exerce dans les domaines suivants :

- établissement des programmes d'enquêtes;
- visa des questionnaires d'enquêtes utilisés par les diverses administrations;
- élaboration des codes et nomenclatures (activités économiques, catégories socio-professionnelles, géographiques, etc.);
- examen des problèmes statistiques des différentes administrations;
- tenue de fichiers centraux.

Cette action s'est également affirmée dans le cadre de la préparation des V^e et VI^e Plans (groupe de travail statistique de la Commission de l'économie générale et du financement, Commission de l'information économique).

3. La participation de l'I. N. S. E. E. à la collecte et au traitement de l'information de base demeure néanmoins très importante :

a) Grands recensements (voir ci-dessous § C 1).

1. On évoquera en passant la querelle entre la centralisation au sein d'un office ou d'un institut central unique de la fonction de collecte ou au contraire de la décentralisation au bénéfice des ministères techniques de tutelle. Nous pensons que la seconde solution est préférable moyennant le respect de certaines conditions fondamentales :

- identité de salaires quelle que soit l'affectation;
- position hiérarchique satisfaisante des services;
- neutralité des services techniques à l'égard des services statistiques et des résultats numériques diffusés;
- respect du secret statistique;
- intérêt des responsables politiques du développement de ces services et de l'adéquation des programmes aux besoins.

2. La loi du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques a donné réglementairement ce rôle à l'I. N. S. E. E.

Département ministériel	Cellule statistique	Compétence	Date de création C de 1 ^{re} cellule	Observations
Ministère d'État chargé des Affaires sociales	Division de la statistique et des études	Travail	1964	
Ministère de la Santé publique	Division des études et du Plan	Santé	1961	
Ministère du Développement industriel et scientifique	Service central de la statistique et des informations industrielles	Industrie	1947	Avant cette date existence de sections techniques non coordonnées ⁽¹⁾
Ministère de l'Agriculture et du Développement rural	Service central des enquêtes et études statistiques	Agriculture	1945 puis 1957	Échelons régionaux et départementaux rattachés directement à la cellule centrale
Ministère de l'Éducation nationale	Service central des statistiques et sondages	Éducation nationale	1965	Échelons régionaux non rattachés directement à la cellule centrale
Ministère de l'Aménagement du territoire, de l'Équipement, du Logement et du Tourisme	Service des statistiques et études économiques	Équipement logement	1966	Échelons régionaux rattachés directement à la cellule centrale
Ministère des Transports	Bureau central des statistiques de transports	Transports	1961	
Ministère de l'Aménagement du territoire, de l'Équipement, du Logement et du Tourisme	Division de la statistique, des études et de la conjoncture	Tourisme	1963	

1. Coordination de 1947 à 1967, en 1967 création d'un Service central procédant à compter de novembre 1970 à des enquêtes directes

b) Enquêtes auprès des ménages (consommation, emploi, intentions d'achat, équipement, transports, logements, etc ⁽¹⁾).

c) Observations des prix (notamment prix de détail).

d) Enquêtes auprès des entreprises : l'I. N. S. E. E. a mis en place des techniques d'observation mensuelle de la conjoncture auprès des industriels, commerçants et artisans sur diverses données (stocks, carnets de commande, perspectives). Dans le cadre de la Commission économique européenne l'I. N. S. E. E. réalise deux enquêtes particulières : structure des salaires et coûts salariaux.

e) Exploitation statistique de documents administratifs.

f) Outre certains travaux des ministères techniques, l'I. N. S. E. E. dépouille :

- les bulletins d'état civil;
- les déclarations annuelles sur les salaires versés par les employeurs;
- l'enquête sur la structure des emplois;
- les bordereaux de cotisation aux Unions de recouvrement de Sécurité sociale et d'Allocation familiales.

g) Gestion de fichiers :

- répertoire d'identification : étaient inscrites sur les registres (par commune) toutes les personnes nées en France depuis 1881. Ces répertoires sont à la base du projet SAFARI (voir ci-dessous § C 3. 2 a);

- fichier électoral;

- fichier des entreprises et établissements industriels et commerciaux. Sur ces registres confiés à l'I. N. S. E. E. par le décret du 15 juillet 1948 figurent chaque entreprise et établis-

1. En 1973 non comprises trois enquêtes pilotes, huit groupes d'enquêtes ont été menées auprès des ménages (200 340 interviews) :

	Ménages
— deux enquêtes de périodicité mensuelle	18 140
— une enquête de périodicité quadrimestrielle	30 000
— une enquête de périodicité semestrielle	10 000
— quatre enquêtes de périodicité annuelle	142 000

sement dont le siège se trouve en France avec son activité principale et sa localisation. Le projet SIRENE (voir ci-dessous § C 3. 2 b) en constitue le développement;

— fichier des véhicules utilitaires.

4. De cette présentation sommaire, on s'aperçoit que l'évolution doctrinale a été progressive au fil des ans, mais qu'en matière de production d'information statistique, deux orientations principales se sont dégagées :

— décentraliser partiellement la fonction de collecte. Cette action a été caractérisée par un certain empirisme;

— utiliser le matériel numérique issu de la gestion administrative des divers ministères.

Cette décentralisation a entraîné :

— l'implantation des cellules extérieures déjà énumérées;

— le renforcement du rôle de coordination de l'I. N. S. E. E. (efforts particuliers sur les nomenclatures et les fichiers, homogénéité des méthodes, amélioration de la qualité des relevés, confrontation des sources).

Le traitement informatique de son côté reste très largement assuré par l'I. N. S. E. E.; néanmoins, il exige des centres de traitement dans les ministères techniques (notamment au ministère de l'Agriculture) pour exploiter les documents d'enquêtes rassemblés par ces ministères et un centre de traitement à Paris, commun à l'I. N. S. E. E. et au ministère du Développement industriel et scientifique (M. D. I. S.).

5. *L'établissement des programmes d'enquêtes.* — Ces programmes dépendent de plusieurs facteurs : outre la contrainte budgétaire évidente, les programmes d'enquêtes doivent s'inscrire dans un ensemble pluri-annuel. Les grandes priorités en ont été fixées pour la période actuelle par la Commission de l'information économique du VI^e Plan 1971-1975 (voir l'annexe 2 du rapport de cette Commission : liste des opérations statistiques prioritaires durant le VI^e Plan). Par ailleurs, par la voie de règlements ou de directives, certaines enquêtes sont dites « communautaires », en ce sens qu'elles rentrent dans le cadre du programme de l'Office statistique des Communautés européennes.

Le programme annuel d'enquêtes est soumis au Comité de coordination des enquêtes statistiques dont chaque section spécialisée (Agriculture, Industrie, etc.) examine les propositions qui la concernent.

Enfin le programme d'enquêtes, tant pour l'I. N. S. E. E. que pour les ministères techniques, doit rentrer dans un cadre trimestriel de réalisation, de telle manière que les moyens (crédits d'enquêtes ou de vacations, structure du réseau d'enquêteurs, charges en analyse et programmation informatique, potentiel heures machines des ordinateurs) soient cohérents avec les charges résultant des réalisations projetées.

En 1972, une nouvelle instance a été créée : le Conseil national de la statistique dont l'un des buts sera d'adapter mieux encore les programmes de travail aux besoins des utilisateurs. Ce conseil suivra l'exécution du programme statistique et émettra des avis dans un rapport annuel.

B — LES FACTEURS RÉCENTS QUI ONT INFLUENCÉ LA NATURE DE LA COLLECTE

Différentes possibilités sont offertes (recensements, enquêtes, fichiers, etc.) lorsque l'on se préoccupe de collecte de l'information.

1. Cette décomposition revêt un aspect de plus en plus formel, car la première des préoccupations présentes est de rechercher la possibilité *de relier entre elles les informations concernant un même secteur statistique*, de manière à faciliter leur enrichissement réciproque ⁽¹⁾. En effet, le développement de l'outil informatique a permis d'envisager des opérations qui auraient semblé antérieurement impossibles par la voie manuelle ou la voie de la mécanographie classique. La fusion de fichiers issus de sources différentes, mais concernant en tout ou partie les mêmes unités statistiques, a paru constituer une voie très riche d'informations nouvelles pouvant être obtenues à un coût raisonnable.

Elle permet, en outre, des contrôles et des vérifications qui sans cela seraient impossibles. Cette technique permettra en particulier un progrès important de la qualité des statistiques sur les entreprises.

2. Le second facteur original a été la recherche *d'informations statistiques à partir du dépouillement d'archives administratives* tenues par les ministères. On pouvait regretter antérieurement que les documents réglementaires établis à l'occasion de l'ouverture d'un dossier, de la liquidation d'une dépense, de la perception d'un impôt ne fassent pas l'objet de traitements statistiques ou de rapprochements avec des sources statistiques proprement dites. Par ailleurs, les enquêtes statistiques recherchent parfois des informations voisines auprès des mêmes agents économiques, on pouvait donc espérer réduire la charge reposant sur les ménages ou les entreprises. Il ne faut d'ailleurs pas se dissimuler que les administrations détentrices de ces archives ne sont pas toujours favorables à ces exploitations pour des raisons diverses :

- désintérêt quant à une utilisation extérieure à leurs préoccupations courantes;
- souhait de « conserver » l'information;
- qualité douteuse des données recueillies.

Quoi qu'il en soit, en dépit des réserves ou des critiques que l'on décrira plus loin, ces dépouillements ont été entrepris depuis plusieurs années et couvrent des domaines assez divers (Bénéfices industriels et commerciaux. Salaires versés, Mutualité sociale agricole, Cotisations versées à la Sécurité sociale, etc.).

3. Le troisième facteur, très différent dans son essence est celui d'une *diversification des données sur le plan géographique*. Autrement dit, le statisticien a subi une pression de plus en plus forte pour que les données nationales soient régionalisées, les décisions de localisation (équipements, logements, entreprises) exigeant de plus en plus d'être éclairées par des statistiques originales.

4. Le développement des activités tertiaires (administration, gestion, enseignement, etc.) constitue un facteur favorable de la statistique (mais cette croyance est parfois aveugle) et comme cette dernière est fréquemment mal armée pour saisir les variations de structures, les demandeurs (cédant d'ailleurs inconsciemment à une inclination naturelle) insistent pour disposer d'informations plus fines sur les structures (détail plus poussé par produit, niveau hiérarchique plus détaillé).

Les dix dernières années ont été marquées par un accroissement très important de la demande d'information au niveau local, ce besoin motivé par de multiples travaux (planification régionale, aménagement rural, implantation d'investissements collectifs, croissance urbaine) a été en partie satisfait par la systématisation de certains dépouillements de gestion administrative et la mobilisation du potentiel informatique, mais certaines techniques de collecte (enquêtes par sondage), répondent mal, par nature, aux besoins de connais-

1. Dans les domaines de l'emploi et de l'industrie par exemple.

sance localisée. On est donc amené en matière de collecte de l'information régionale à plusieurs solutions qui se situent dans le cadre d'un développement de l'information localisée à l'intérieur d'une organisation nationale de la collecte, complétée éventuellement par des collectes régionales coordonnées :

- dépouillements complémentaires de traitements informatiques nationaux;
- étoffement de l'échantillon national afin d'obtenir une meilleure représentativité au niveau régional;
- ajout de questions ou questionnaires complémentaires pour répondre à des problèmes locaux;
- modification de l'unité statistique tabulée (information sur l'établissement et non sur l'entreprise par exemple);
- lancement d'enquêtes spécifiques (exode rural, demande solvable de logements, etc.).

Les avantages de cette promotion de la collecte d'informations régionales sont évidents, mais à cette demande particulière correspond le coût d'une production statistique spécifique :

- augmentation des coûts de collecte;
- alourdissement des plans d'échantillonnage;
- dispersion des travaux et risque de « concurrence » entre travaux nationaux et régionaux;
- complication des chaînes de traitement informatique, si l'on n'est pas attentif à la coordination des cadres de présentation et des nomenclatures.

Néanmoins, et du fait même :

- de la décentralisation de l'I. N. S. E. E. par l'implantation de ses unités régionales;
- de l'intention affirmée par les divers gouvernements de voir confier aux autorités régionales certaines responsabilités politiques, économiques et administratives, on peut penser que l'évolution naturelle conduira à un poids de plus en plus sensible des actions statistiques régionales ⁽¹⁾.

C — LES PROCÉDURES DE COLLECTE ET LES PROBLÈMES QU'ELLES POSENT

On examinera successivement les recensements, les enquêtes par sondage et les systèmes de référence administratifs.

1. *Les recensements.* Un effort non négligeable a été réalisé depuis 1945 et des recensements généraux ont été lancés dans divers domaines :

- Recensement de la population et des habitations : 1946, 1954, 1962, 1968.
- Recensement de l'agriculture : 1954-1955, 1970-1971.
(p. m. enquête communautaire sur les structures agricoles : 1967-1968).
- Recensement de l'industrie : 1963 (sur exercice 1962).
- Recensement des transports : 1963.
- Recensement de la distribution ⁽²⁾ : 1967 (sur exercice 1966).
- Dénombrement des services : 1971 (sur activités 1970).

1. P. CARRÈRE, le développement de l'information économique régionale en France. Conférence devant la Commission mixte de coordination statistique, Barcelone, avril 1970.

2. Utilisation de deux types de questionnaires, opération exhaustive pour les entreprises de 3 salariés et plus, échantillonnage pour les autres.

Un fait nouveau apparaît : la remise en question (totale ou partielle) de certaines opérations (agriculture notamment) en raison de leur lourdeur sur le terrain, du coût et des délais du dépouillement. Dans le domaine industriel, la combinaison des enquêtes annuelles de secteurs et des enquêtes de branches pourrait rendre moins nécessaire la réalisation d'un recensement industriel.

Diverses solutions ont été adoptées :

- Dépouillement exhaustif léger et dépouillement lourd par sondage (R. P., 1968).
- Dépouillement exhaustif préliminaire sur une partie du questionnaire (R. G. A., 1970-1971).

Néanmoins à l'horizon 1980, on peut se poser la question de savoir si un recensement général de l'agriculture ne devrait pas comporter deux types de questionnaires l'un léger pour les exploitations « économiquement non intéressantes », l'autre plus ambitieux pour les « vraies exploitations ». Ceci suppose la disponibilité d'un fichier à jour et détaillé en vue de la stratification envisagée. Cette solution n'est pas nouvelle puisqu'elle est adoptée aux U. S. A. en prenant comme critère séparateur le chiffre d'affaires de l'exploitation agricole ⁽¹⁾.

Dans le domaine des recensements de population, on pourrait envisager comme en ALLEMAGNE ⁽²⁾, l'utilisation d'un bulletin sommaire, sauf pour un échantillon de logements au sein duquel des questionnaires plus complets seraient remplis.

Plus généralement deux remarques sont à noter :

a) Le sondage peut intervenir au stade de la collecte ou du dépouillement.

Au stade de la collecte, deux méthodes peuvent être préconisées :

- dénombrement exhaustif, interrogation plus complète d'un échantillon;
- dénombrement par sondage.

Dans ces cas, le choix des unités à interroger pose des problèmes techniques (risque accru de dénombrements incomplets par exemple) ou psychologiques (dans une opération de grande envergure, le principe du sondage heurte l'idée d'égalité des devoirs du citoyen; la même raison conduit à ne pas faire d'élections par sondage).

b) Que le sondage intervienne au stade de la collecte ou du dépouillement, la méthode entre en conflit avec les objectifs régionaux et locaux des grands recensements. Ainsi, par exemple, on a dû renoncer, pour le dépouillement par sondage du recensement de 1968, à certains contrôles de cohérence effectués lors du recensement de 1962 afin de réduire l'erreur aléatoire (tirage de logements et non d'immeubles si bien que les fichiers logements et immeubles ne sont plus parfaitement cohérents).

2. Les enquêtes par sondage

2.1. Généralités

Après un développement relativement limité des enquêtes par sondage pendant la période 1945-1960, cette technique d'information s'est répandue à un rythme accéléré pendant la période 1960-1972 sous la pression d'une demande très forte et parfois insuffisamment programmée.

1. C. TAUBER, American Journal of Agricultural Economics : les données des futurs recensements de l'agriculture, mai 1972.

2. Recensement de la population en Allemagne, *Économie et Statistique*, novembre 1970.

La phase de scepticisme du public et de l'Administration à l'égard de cette technique a été suivie d'une période où une confiance parfois aveugle a conduit le statisticien à s'éloigner parfois un peu trop des possibilités normales de cette méthode d'investigation.

Les buts actuels des enquêtes par sondage sont beaucoup plus difficiles à atteindre : il est infiniment plus délicat de connaître des variations de structure ou de comportement des agents économiques que d'appréhender cette structure elle-même (ex. variations des dépenses des ménages plutôt que niveau des dépenses). Sous l'influence d'une demande exubérante, des problèmes très techniques ont été parfois abordés dans des enquêtes auprès des ménages (ex. : renseignements sur les garanties des prêts au logement...), les sujets traités dans une même enquête se sont multipliés (ou ce qui revient au même, on a regroupé deux enquêtes distinctes de façon à réduire les coûts de collecte).

Alors que les préoccupations des demandeurs étaient souvent orientées à court terme ou voulaient lier la connaissance statistique aux dates des grandes décisions politiques (Plan notamment), les services statistiques (et surtout l'I. N. S. E. E.) arrivaient difficilement à faire face à cette demande foisonnante et aux nouvelles méthodes d'exploitation (notamment informatique); tandis que l'effort des statisticiens s'est porté davantage vers l'analyse que vers la collecte, la croissance trop rapide du volume des enquêtes a conduit à une crise qui s'est traduite par un allongement des délais d'exploitation ⁽¹⁾ et une détérioration progressive de la qualité.

Malgré ces difficultés, qui ne sont pas encore aplanies, on peut distinguer quelques lignes directrices dans le plan de développement statistique de l'Administration française; trois caractéristiques essentielles ont marqué le programme d'enquêtes français :

- un programme d'enquêtes cohérent;
- un souci de l'aléatoire;
- la recherche de la validation.

2.2. *Un programme d'enquêtes cohérent*

A peu près au même moment, se sont développées les enquêtes agricoles et les enquêtes auprès des ménages; le programme d'enquêtes industrielles ayant été développé plus tard...

Le programme des enquêtes agricoles a été basé sur la réalisation d'enquêtes de structures (dites horizontales) sur lesquelles venaient se greffer des enquêtes spécifiques (verticales) aux objectifs plus fins et plus prévisionnels. Des essais d'enquêtes prévisionnelles ont été réalisés dans le domaine du cheptel laitier et de la production laitière.

Le programme d'enquêtes auprès des ménages a été axé essentiellement sur la recherche d'informations économiques factuelles : les particuliers en tant que producteurs (enquêtes sur l'emploi), en tant que consommateurs (enquêtes de consommation, enquêtes sur l'épargne); les autres thèmes d'enquêtes (sociologiques, utilisation des loisirs...) commencent seulement à se développer à l'I. N. S. E. E.

L'axe central du programme auprès des ménages a été constitué par quatre enquêtes périodiques ou permanentes :

- l'enquête annuelle sur l'emploi, destinée à saisir les variations à moyen terme de la population active, répartie par grandes activités économiques, dont l'échantillon est renouvelé par parties;

1. Pour des enquêtes légères, des programmes ne comportant pas de redressement et limitant les critères de répartition permettent de disposer rapidement des tableaux statistiques nécessaires.

— deux enquêtes continues sur les budgets familiaux et la consommation alimentaire dont les échantillons sont constamment renouvelés;

— une enquête périodique de conjoncture sur les intentions d'achat des consommateurs dont l'échantillon est renouvelé par parties.

A cet axe central, s'ajoutent des enquêtes occasionnelles destinées à approfondir les thèmes abordés par ailleurs (formation professionnelle, domaines particuliers de la consommation...), ou à les compléter par des thèmes voisins (épargne, emploi du temps, loisirs...).

La permanence de l'axe central devait permettre de faciliter la comparabilité des enquêtes; l'expérience montre, en effet, que c'est au prix d'une répétition à l'identique que l'on peut assurer la comparabilité des résultats des enquêtes — et encore, on le verra plus loin, cette condition nécessaire n'est pas suffisante. On remarque d'ailleurs que la fiabilité des enquêtes spécialisées est parfois supérieure à celle des enquêtes à objectifs plus diversifiés.

2.3. *Le souci de l'aléatoire*

La plupart des enquêtes par sondage effectuées par le secteur public se font auprès d'un échantillon aléatoire; lorsque le thème choisi est peu adopté ou mal accueilli par certaines catégories de la population, on préfère limiter le champ de l'enquête plutôt que de renoncer à la méthode aléatoire (ex. : enquête sur l'épargne limitée aux salariés et inactifs, enquête sur les budgets temps limitée aux personnes de moins de 65 ans...).

Un grand nombre d'enquêtes ont un plan de sondage classique stratifié selon la région et l'importance de l'agglomération, à un, deux ou trois degrés selon la strate. Les particularités des plans de sondage de l'I. N. S. E. E. sont de plusieurs ordres :

- fixité des unités primaires pour de nombreuses enquêtes et parfois des ménages;
- problèmes liés à l'absence de mise à jour régulière de la base de sondage « ménages »;
- multiplicité des unités statistiques et choix de la méthodologie des enquêtes sur des budgets familiaux;
- tentatives pour « redresser » les échantillons, déformés du fait des refus ou des absences;
- emploi de plans de sondage particuliers pour certaines enquêtes spécifiques (échantillon d'aires dans l'enquête sur l'emploi, photographies aériennes dans les enquêtes agricoles, etc.).

2.3.1. *Fixité des unités primaires et parfois des unités de sondage*

Les avantages de cette fixité sont d'abord d'ordre pratique :

- emploi d'un réseau d'enquêteurs décentralisé, bien implanté dans les lieux d'enquête, et conservé pourtant pour plusieurs années; sa formation est mieux rentabilisée. Il ne faut pas confondre cependant la fixité de l'échantillon maître et la permanence d'un réseau d'enquêteurs permanents;
- limitation du coût de constitution de la base de sondage (établissement de la liste des unités de sondage uniquement pour les unités primaires-échantillon, ce qui représente un avantage tout particulier pour les listes de logements récents);
- possibilité d'étudier l'évolution chronologique d'un échantillon fixe d'unités et traitement statistique des évolutions par matrices de passage;
- gain de précision dans la mesure des variations temporelles (surtout si, comme dans l'enquête sur le logement de 1970, on utilise à nouveau le même échantillon de ménages que dans l'enquête précédente);

- possibilité d'utilisation de données tirées d'une enquête pour compléter l'échantillon d'une autre enquête (ainsi l'enquête sur l'emploi de 1967 a permis d'augmenter la taille de l'échantillon de ménages de l'enquête sur les logements récemment emménagés sans imposer de déplacements considérables aux enquêteurs);

- réduction légère de l'importance des ménages éliminés de l'enquête, car difficiles à contacter (l'enquêteur résidant à proximité).

En revanche, cette procédure peut avoir des inconvénients :

- la distance entre les enquêteurs et l'Administration est plus grande que si les enquêteurs résidaient près des agents chargés de leur encadrement; leur formation et le contrôle de leur travail sont trop souvent insuffisants;

- dans certaines zones, il peut exister entre l'Administration locale et l'enquêteur des liens préjudiciables au travail d'enquête (par exemple, dans une zone où étaient relevés régulièrement des prix, l'enquêteur était trop bien introduit auprès de la mairie; pour supprimer cet inconvénient on a dû changer la zone échantillon);

- la connaissance mutuelle que peuvent avoir entre eux les enquêtés peut conduire à des réactions collectives : refus en chaîne, ou, ce qui est pire, acceptation partielle, mais méfiante à l'égard de cette personne connue qu'est l'enquêteur;

- les études en cours semblent faire ressortir certaines critiques à l'encontre des enquêteurs permanents.

2.3.2. *Problèmes liés à l'absence de mise à jour régulière de la base de sondage « ménages »*

Il n'existe pas en France de fichier de population tenu à jour où figure l'adresse des individus.

L'I. N. S. E. E. dispose en revanche des listes de logements du recensement précédent; ce sont ces listes qui servent de base de sondage; les unités de sondage sont, en général, les logements; on appelle ménage l'ensemble des habitants d'un même logement; c'est le ménage occupant au moment de l'enquête le logement désigné par le sort qui est interrogé.

Cette procédure a des avantages — utilisation des informations tirées du recensement pour stratifier, contrôler, redresser l'échantillon, — mais elle a aussi des inconvénients.

Tout d'abord, elle est liée à la qualité du dénombrement effectué au moment du recensement : les omissions et les doubles comptes du recensement sont aussi ceux des enquêtes. De plus, pour que tout ménage habitant au moment de l'enquête ait une probabilité connue d'être tiré au sort, il faut que l'échantillon contienne également des locaux ayant eu d'autres affectations au recensement (logements vacants, résidences secondaires); ces compléments nécessaires sont coûteux (de 8 à 15 % des échantillons) et parfois sources d'erreurs (confusions entre résidences secondaires et résidences principales, entre absences temporaires et résidences secondaires).

Ensuite, la délimitation du local est source d'erreurs; cette raison est à l'origine du recours à un échantillon aréolaire pour l'enquête sur l'emploi (*cf.* ci-dessous); en effet, malgré toutes les précautions prises, des personnes habitant des parties du local relativement indépendantes (chambres de personnel par exemple) échappaient à l'enquête.

Mais surtout, il convient de compléter la base de sondage par des listes de logements achevés depuis le recensement. La qualité de cette base de sondage complémentaire a été toujours à l'origine de problèmes imparfaitement résolus : faute de dénombrement des achevements, on utilise des autorisations de construire; pour l'échantillon d'unités primaires, on vérifie deux fois par an l'état d'avancement des bâtiments et on retient les logements

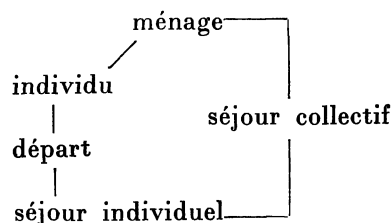
correspondants quand ils sont habités. Le coût de cette mise à jour est élevé mais surtout l'utilisation de deux bases de sondage distinctes a des inconvénients importants :

- raccordement imparfait des deux bases de sondage;
- caractère incomplet de la base de sondage complémentaire (certains logements allant des cabanes de bidonvilles jusqu'à certaines résidences secondaires de luxe sont construits sans autorisation);
- erreurs matérielles (une construction peut faire l'objet de plusieurs autorisations, surtout si le projet initial est modifié; inversement, la transmission de certains documents est imparfaite);
- mise à jour avec retard (ce qui peut expliquer des biais dans l'estimation de certaines dépenses d'habitation : mobilier par exemple...);
- augmentation considérable de l'erreur aléatoire sur les effectifs de ménages, du fait du développement des villes nouvelles (d'où la nécessité pratique de distinguer le plan de sondage des logements tirés du recensement de celui des logements neufs, surtout pour les constructions importantes).

Un fichier informatisé complet des logements neufs (SIROCO) est en cours de constitution; la validité de la base de sondage de logements neufs dépendra du sérieux avec lequel ce fichier sera tenu à jour par les Services de l'équipement.

2.3.3. *Multipllicité des unités statistiques, méthodologie des enquêtes sur les budgets familiaux*

Dans la plupart des enquêtes des unités statistiques auxiliaires, de différents niveaux, sont analysées; l'unité de sondage correspond en général à une grappe d'unités statistiques qui sont toutes étudiées. Ainsi, par exemple, les enquêtes sur les vacances fournissent des renseignements pour chaque ménage (unité de sondage) sur les individus qui le composent (unité de niveau 2), sur leurs départs en vacances (unité de niveau 3), sur leurs séjours individuels (unité de niveau 4 — un départ peut correspondre à plusieurs séjours successifs); enfin, pour l'étude des dépenses de vacances, on peut être conduit à regrouper des séjours individuels et à définir des séjours collectifs (unité de niveau 2).



Parfois, on ne retient pas la totalité de la grappe d'unités statistiques correspondant à l'unité de sondage : lorsque l'indépendance des réponses des membres d'un même foyer serait douteuse, on en tire un, par exemple, par la méthode de Kish (enquête sur les budgets temps par exemple).

Le problème de l'exécution des enquêtes sur les budgets familiaux a été posé jusqu'ici en termes analogues; la méthodologie employée consistait à confier à l'enquêteur pendant une courte période un grand nombre de ménages tirés au sort; ainsi les carnets de comptes qui sont complétés par des interviews sur les dépenses rares ou contractuelles sont remplis seulement pendant une semaine; sur le plan de la précision de la mesure, cette solution est certainement préférable à celle qui consisterait à soumettre à l'enquête un nombre

beaucoup moins grand de ménages pendant une période plus longue, un mois, par exemple : un ménage-mois peut être considéré comme une grappe de 4 ménages-semaine, grappe qui n'est efficace que pour des dépenses relativement rares ou périodiques. Nous verrons plus loin que d'autres considérations sur la qualité des données recueillies nous conduisent à remettre en cause cette pratique.

Dans d'autres cas, l'unité de sondage peut être même plus petite que l'unité statistique étudiée (voir ci-dessous 2.4.2.).

2.3.4. Tentatives pour « redresser » les échantillons déformés du fait des refus ou des absences de longue durée

Dans ses enquêtes auprès des ménages ⁽¹⁾, l'I. N. S. E. E. ne fait pas appel à la loi sur l'obligation statistique de répondre, préférant une collaboration sincère des enquêtés à une contrainte peu efficace.

Rappelons également qu'aucun intermédiaire n'est utilisé pour faciliter la coopération des ménages ⁽²⁾.

Ainsi, une proportion assez considérable des ménages soumis aux enquêtes refusent de répondre ou ne peuvent être joints par l'enquêteur (absents de longue durée). Cette proportion dépend de la difficulté de l'enquête; elle a atteint 33 % dans l'enquête sur les budgets familiaux effectuée dans le cadre de l'O. S. C. E. en 1963-1964, 12 % seulement dans le cas d'enquêtes plus faciles, à une seule visite.

Les ménages non répondants ont des caractéristiques moyennes particulières :

- plus urbains que ruraux;
- correspondant plus souvent à des isolés qu'à des familles nombreuses;
- très aisés ou au contraire particulièrement modestes;
- étrangers plutôt que français.

Le redressement de l'échantillon consiste à traiter l'échantillon de répondants comme un échantillon par quota. L'idée directrice, parfois contestée, est que les ménages non répondants d'une certaine strate définie par des caractéristiques socio-démographiques connues sont plus proches des ménages répondants de la même strate que l'ensemble des ménages répondants ⁽³⁾.

On peut effectivement se demander lorsque le taux de non réponse devient très important si l'enquête ne se rapproche pas d'une enquête par quota et s'il ne vaut pas mieux renoncer totalement à l'aléatoire et se montrer plus exigeant dans la qualité de la collecte.

En pratique, c'est la solution qui a été retenue pour le réseau d'information comptable agricole, le tirage aléatoire d'un échantillon d'exploitants s'étant révélé impossible.

2.4. Emploi de plans de sondages particuliers

2.4.1. Échantillon d'aires géographiques

Ce schéma facilite l'implantation d'enquêteurs semi-permanents, réduit les coûts de déplacements, mais peut majorer l'erreur de sondage par l'effet de grappe. Un bon exemple est constitué par l'enquête sur l'emploi ⁽⁴⁾.

1. Il en est de même des enquêtes sur les prix effectuées auprès des commerçants.

2. Dans les enquêtes effectuées dans la plupart des pays de l'Est les personnes soumises à l'enquête sont tirées au sort par l'intermédiaire de l'entreprise. La même procédure a été adoptée pour plusieurs enquêtes effectuées dans certains milieux particuliers en France (enquête sur les budgets familiaux de la C. E. C. A., 1956-1957, enquête sur les budgets familiaux effectué par le CREDOC à la demande de l'UNCAF — Union des Caisses d'Allocations familiales).

3. Elle est évidemment contestable dans certains cas : ménages refusant de répondre à une enquête sur les dépenses de santé, car ayant des malades dans le foyer, par exemple.

4. R. POHL, P. LAULHE, R. LIONNET, enquête sur l'emploi, 1971, I. N. S. E. E., Collection D-12.

Après stratification de « segments » de 120 logements contigus (en milieu rural) ou de 40 logements (en milieu urbain), un échantillon est tiré systématiquement à équiprobabilité. Cet échantillon est renouvelé annuellement par tiers parmi les 1 557 aires initiales.

2.4.2. *Utilisation d'unités de sondage auxiliaires*

Par ce terme, on entend l'utilisation d'une base de sondage qui ne fournit pas directement les unités à échantillonner.

On peut rappeler l'exemple de l'enquête sur la mécanisation de l'agriculture (1962-1963) : disposant d'un fichier de tracteurs agricoles immatriculés au 1/1/1962 stratifié par région agricole et selon l'année de la dernière immatriculation, on tirait un échantillon des possesseurs de tracteurs. Il y avait donc, moyennant pondération adéquate des résultats, passage d'un fichier de tracteurs aux exploitations mécanisées (unités d'observation).

Un autre exemple est constitué par les enquêtes sur le rendement des cultures : on tire un échantillon de points sur des photographies aériennes pour lequel, au cours d'un premier passage d'enquêteurs, on a relevé la nature de la culture. Ces points engendrent des champs sur lesquels on procède à deux coupes échantillons.

2.4.3. *L'utilisation de la photographie aérienne*

Cette utilisation s'est largement développée ⁽¹⁾ et ses applications sont multiples :

- utilisation des terres (par visite sur le terrain par des enquêteurs de points implantés systématiquement sur un échantillon de photographies);

- déroulement de la campagne agricole (par visite sur le terrain à différentes époques de l'année de points situés en dehors des prairies permanentes ou de forêts);

- tirage d'exploitations ayant une spéculation donnée par « remontée » de la parcelle à l'exploitant : au cours d'une enquête sur la structure des vergers, il a été étudié 9 440 points sur le terrain à partir desquels on a enquêté auprès des exploitants concernés;

- contrôle de la qualité d'un recensement général de l'agriculture par tirage d'un échantillon de points, et vérification de l'existence de l'exploitation et de la mention sur le questionnaire de la culture repérée au sol;

- utilisation de la photographie aérienne pour la constitution d'un schéma d'îlots et d'une maille de voies permettant ultérieurement la codification automatique (dans le cadre d'un recensement de population) au lieu de résidence et du lieu de travail;

- utilisation de la photographie aérienne pour la constitution d'aires échantillons au sein desquelles certaines opérations de contrôle d'une opération antérieure peuvent être exécutées. C'est selon cette procédure qu'on a vérifié en France, pour le recensement de la population de 1946, l'exhaustivité du dénombrement des entreprises industrielles et commerciales ⁽²⁾.

2.4.4. *Les enquêtes à plusieurs passages*

Cette technique résulte de l'impossibilité dans certains cas d'obtenir des renseignements rétrospectifs pour une période antérieure. On doit alors se résoudre à étudier soit un échantillon permanent soit un échantillon renouvelé sur lesquels on procède aux observations répétées nécessaires.

1. Étude sur l'utilisation du territoire. Méthodologie. Résultats, 1969-1970-1971, S. C. E. E. S., supplément série *Études*, n° 104.

2. G. CHEVRY, Control of a general census by means of an area sampling method, *Journal American Statistical Association*, septembre 1949.

On peut citer :

— Les enquêtes démographiques à plusieurs passages dans les pays africains et malgache d'expression française qui fournissent des informations sur certains événements démographiques (naissances, décès, migrations) ⁽¹⁾. Le rythme des visites peut être mensuel ou tous les deux mois;

— les enquêtes sur les conditions de vie et la consommation alimentaire des ménages ⁽²⁾;

— certaines enquêtes sur le rendement des cultures pour lesquelles il y a récolte continue et non récolte effectuée en une seule opération ⁽³⁾. Le nombre et la périodicité de ces passages sont liés à la nature du produit et à la coopération des agriculteurs. Certains de ces passages peuvent constituer d'ailleurs en de simples visites, et non des coupes échantillons, lorsque le produit est encore très éloigné de la maturité ⁽⁴⁾.

2.4.5. *Échantillons composites*

Dans le domaine agricole, il faut signaler une expérience originale actuellement en cours coordonnée au niveau de la Communauté économique européenne et qui a débuté en 1968 : le réseau d'information comptable agricole. Ce réseau comportait 2 580 exploitations agricoles stratifiées selon trois critères (critère géographique — taille de l'exploitation — orientation technico-économique), il est d'environ 3 000 exploitations en 1972 (l'échantillon européen comporte pour les six pays 10 000 exploitations). L'échantillon est constitué d'exploitants volontaires répondant aux conditions de certains quota, qui sont interrogés tous les 2 ou 3 mois par des enquêteurs à formation comptable. Il est demandé aux exploitants de conserver le maximum de documents afin de pouvoir reconstituer leur comptabilité. Les résultats doivent mettre en lumière les disparités entre groupes homogènes et définir des fonctions de production, plutôt que des données substituables à celles de la statistique agricole classique. Compte tenu de l'expérience acquise, à cette étude relativement onéreuse (de l'ordre de 900 000 US \$ par an) pourraient être substituées trois opérations différentes et indépendantes. Pour les petites exploitations : étude d'un échantillon aléatoire avec l'emploi d'un questionnaire statistique léger, pour les exploitations moyennes : utilisation d'une partie des comptabilités déjà existantes dans les centres de gestion et d'économie rurale des organisations professionnelles avec redressement *a posteriori* des résultats convenablement stratifiés, pour les exploitations importantes : utilisation des documents détenus par l'Administration financière (imposition au bénéfice réel). On voit apparaître l'orientation de l'utilisation simultanée de plusieurs approches différentes.

2.5. *La recherche de la validation*

Pour de nombreuses enquêtes, il a paru utile de confronter au niveau individuel les données des enquêtes par sondage avec celles issues de documentations administratives.

L'intérêt est multiple :

— croisements de données d'une documentation par une autre;

1. Les enquêtes démographiques à passages répétés. Application à l'Afrique d'expression française et à Madagascar. Méthodologie, I. N. S. E. E., ORSTOM, INED, 1971. Voir également C. SCOTT et J. B. COOKER, *Sample design in space and in time* et PODLEWSKI, Essai d'observation permanente des faits d'état civil dans l'Adamaonor. Recherche méthodologique, ORSTOM, 1970.

2. R. HALLUS, réalités africaines et enquêtes Budget Consommation, *Économie et Statistique*, avril 1970.

3. C. RIPAILLES et J. ROSSION, Préviation de récolte pour le cacao. *Café-Cacao-Thé*, avril-juin 1968.

4. En matière de prévision de production fruitière, citons à titre illustratif en France pour 1972 l'observation de 2 000 parcelles (pêche, pomme, poire) à quatre reprises pour la pêche et cinq reprises pour les deux autres fruits.

- contrôle de qualité des enquêtes auprès des ménages;
- réalisations de tests de validité d'une technique d'enquête.

Citons par exemple :

a) l'enquête permanente sur l'épargne. Cette enquête est effectuée auprès d'un échantillon de ménages dont on note les numéros de comptes chèques postaux et de livrets de Caisse d'épargne. Ces organismes communiquent à l'I. N. S. E. E. l'état et les flux des comptes correspondants. Les deux sources d'information permettent d'obtenir un enrichissement mutuel;

b) le niveau des loyers des locataires et des mensualités versées par les accédants à la propriété, a paru fréquemment très faible aux utilisateurs des enquêtes. Il n'existe, en dehors des enquêtes auprès des ménages, que peu de renseignements sur les variables; la principale d'entre elles est constituée par les dossiers d'allocations logement des Caisses d'Allocations familiales, aussi, a-t-on confronté pour les allocataires du Régime général de Sécurité sociale les principales variables figurant dans les deux sources : loyers, remboursements de prêts, revenus ⁽¹⁾...

c) les renseignements recueillis dans une enquête sur les dépenses de santé ont été systématiquement confrontés, pour chaque acte médical et chaque produit pharmaceutique acheté pendant la période de l'enquête par un échantillon de ménages, avec les dossiers de remboursement des assurés sociaux du Régime général. Cette opération s'est avérée fort coûteuse, mais fort instructive sur le plan pratique;

d) avant de lancer une enquête sur le prix des logements récemment acquis, on a affecté une opération de contrôle en ajoutant à un échantillon aléatoire de logements neufs un échantillon de logements construits par un organisme constructeur important; la confrontation des données des deux sources a permis de vérifier que, les ménages ne connaissent que très approximativement le prix du logement qu'ils venaient d'acheter;

e) dans le domaine de la statistique agricole, la pratique du recoupement de deux estimateurs ou plus résultant d'enquêtes indépendantes est fréquente;

f) la comparaison au cours de deux enquêtes successives d'une action prévue et de l'opération effectuée : intentions d'achat déclarées et achats ultérieurs réalisés.

3. Les systèmes de référence administratifs

3.1. Généralités

Le développement des installations informatiques ⁽²⁾ au sein des administrations a favorisé le sentiment qu'il y aurait avantage à mobiliser à des fins statistiques des fichiers

1. CREDOC, quelques caractéristiques des bénéficiaires d'une allocation logement au début de 1967, avril 1969.

2. La mise en œuvre de l'outil informatique pose un certain nombre de problèmes techniques sur lesquels on ne pourra s'étendre longuement dans le cadre de la présente note. Il faut rappeler le problème déontologique soulevé par l'usage de ces matériels : les chefs d'entreprises sont l'objet de demandes d'information les plus diverses et il peut être tentant de songer à les en dispenser par la création d'un fonds documentaire auquel les demandeurs feraient appel en cas de besoin. Néanmoins, outre les difficultés méthodologiques qui seront développées ci-dessous, on ne peut négliger celle qui est relative à la protection des libertés individuelles et de la gestion commerciale et industrielle. Si le droit à l'information est nécessaire au fonctionnement démocratique d'une collectivité, la puissance de l'outil ne peut être réservée à quelques-uns et un véritable droit d'usage doit être élaboré. Le statisticien ne peut rester indifférent ou étranger à cette recherche à la fois comme technicien et comme citoyen. Voir à ce sujet Martin Laprade, Développement des statistiques démographiques et sociales et protection de la vie privée. Union internationale pour l'Étude scientifique de la population. Liège, août 1973.

de gestion dans la mesure où leur qualité ne serait pas mise en cause (exhaustivité nationale, mise à jour, fiabilité des données). Cette hypothèque étant levée, on est passé très vite à l'articulation de ces divers fichiers et à la conception *de systèmes d'information*.

Ces systèmes ont sur le plan de la collecte statistique deux finalités :

- unifier les canaux de collecte, c'est-à-dire créer des fichiers de référence;
- permettre la réunion d'informations tirées d'enquêtes statistiques ou d'opérations administratives, les données numériques étant, soit individuelles, soit agrégées ⁽¹⁾, ⁽²⁾.

3.2. Les systèmes de références

Les principaux systèmes de références mis en œuvre ou projetés en France sont les suivants :

a) Pour les personnes :

Système automatisé pour les fichiers administratifs et le répertoire des individus (S. A. F. A. R. I.). A partir des registres d'état civil tenus dans les mairies avaient été constitués par l'I. N. S. E. E., dès 1941, des répertoires de personnes physiques. Chaque Direction régionale gérait les répertoires des personnes nées dans les communes de la circonscription d'action régionale. Les mises à jour étaient effectuées à l'aide des documents transmis par les mairies. Il a été décidé d'automatiser ces répertoires et le nouveau répertoire, une fois sa transcription sur support magnétique effectué, sera exploité en un seul lieu. A chaque personne repérée par son nom, ses prénoms (l'adresse n'est pas mentionnée), son sexe, sa date de naissance et ses identifications géographiques de naissance, est attaché un *numéro national d'identité*.

— Sexe	1
— Année de naissance	2
— Mois de naissance	2
— Département de naissance	2
— Commune de naissance.	3
— Rang d'inscription au répertoire	3
— Clé de contrôle	2 ⁽³⁾

15 chiffres caractéristiques

b) Pour les unités économiques :

b.1. Le système informatique pour le répertoire des entreprises et établissements (S. I. R. E. N. E.) a pour objet de se substituer au répertoire actuel qui attribuait à chaque entreprise ou établissement un numéro d'identité à 12 chiffres :

3 chiffres : activité économique en référence à la nomenclature des activités économiques de 1959;

1. La régionalisation entraîne en raison de la différence des thèmes et des besoins des utilisateurs un besoin d'informations statistiques détaillées pour les zones urbaines. Voir notamment *Système d'information pour la planification de l'aménagement urbain* par Jerzy REGULSKI. Conférence des Statisticiens européens WG : 18-29-9 mars 1971. Cet ensemble de données a une structure multidimensionnelle et le système d'information qui en résulte doit être basé sur la coordination des répertoires ou des fichiers des trois catégories qui vont être décrites ci-dessous.

2. En matière de répertoires centraux d'unités économiques dans certains pays, on se reportera au document E/CN3/414 du 20 mai 1970 de la Commission statistique des Nations Unies : *Central Registers of Economic Units*.

3. La clé de contrôle est le complément à 97 du reste de la division par 97 des treize premiers chiffres.

2 chiffres : département de l'établissement ou du siège de l'entreprise;

3 chiffres : commune de l'établissement ou du siège de l'entreprise;

1 chiffre : rang de l'exploitant (augmente d'une unité à chaque changement d'exploitant);

3 chiffres : numéro d'ordre dans la commune pour une activité donnée.

Ce système présentait divers inconvénients (pas de liaison entre le numéro d'une entreprise et ceux de ses établissements, changements d'activité et de localisation, mise à jour lente, etc.).

Dans le nouveau système, l'identification d'une entreprise (numéro SIREN) se fait à l'aide d'un numéro à 8 chiffres, complété d'une clef de contrôle, qui n'est qu'un numéro d'ordre sans signification. Le numéro d'établissement (n° SIRET) s'obtient en complétant ces neuf chiffres d'un numéro interne de l'établissement dans l'entreprise qui contient de 2 à 5 chiffres (dont une seconde clef de contrôle).

Ces numéros sont naturellement complétés, pour chaque unité, de codes et d'informations permettant d'identifier et de classer l'unité : raison sociale et adresse en clair, codes géographique, postal, code d'activité principale exercée, forme juridique, nature de l'établissement, etc. De plus chaque enregistrement d'établissement contient des informations sur l'ensemble de l'entreprise.

b.2. *Le système unifié de statistiques d'entreprises (S. U. S. E.)* a pour objet de constituer le noyau d'informations issues soit des enquêtes statistiques, soit de fichiers de gestion administrative (ministère des Finances, ministère du Travail, Banque de France, Douanes, etc.). On vise à :

- introduire un identifiant propre à une entreprise donnée quel que soit le fichier ou l'enquête où elle est repérée. Cet identifiant sera le numéro S. I. R. E. N. E.;
- homogénéiser les nomenclatures (activités économiques, produits, plan comptable);
- harmoniser les définitions des variables;
- rendre identique les valeurs attribuées dans plusieurs fichiers à une même variable caractéristique d'une même unité de base.

En matière d'enquêtes sur les entreprises, il faut en effet signaler trois préoccupations. La première est relative à *l'analyse des structures* : dépassant le tableau des échanges industriels qui, dans les années 50, constituait l'instrument privilégié de synthèse, sous la pression de divers besoins (chômage, emploi, concentration des unités), sont apparues les recherches sur des groupes d'unités de production, les problèmes de sous-traitance, la répartition des investissements, etc. C'est notamment de ce besoin qu'est née l'idée de la constitution d'un système articulé d'informations sur les entreprises.

La seconde préoccupation est *de suivre dans le temps un échantillon fixe d'entreprises* ⁽¹⁾, et d'étudier le comportement des entreprises, ceci entraîne alors des recherches sur :

- les typologies permettant les analyses structurelles ⁽²⁾;
- les modalités pratiques de suivi de fichiers;
- les méthodes d'analyse sur la morphologie d'un échantillon d'entreprises.

1. La même préoccupation est apparue dans le secteur agricole pour les exploitations, voir : a) problèmes posés par la mise en place entre deux R. G. A. d'un échantillon d'exploitations agricoles (cahiers mensuels de statistique, octobre-novembre 1969); b) étude statistique de l'évolution d'un groupe de 350 coopératives agricoles les plus importantes pour la période 1965-1970, S. C. E. E. S., ministère de l'Agriculture, 1972.

2. Établissement d'une typologie objective des exploitations agricoles françaises, supplément série *Études*, n° 116, S. C. E. E. S., juillet 1973.

Le matériel numérique est issu comme déjà indiqué :

- soit de fichiers administratifs;
- soit de source de gestion ou de réglementation;
- soit d'enquêtes spécifiques (dans ce dernier cas, ces enquêtes sont généralement dans le secteur industriel réalisées par correspondance, des relances pouvant être éventuellement exécutées par des enquêteurs spécialisés). Très souvent, par souci d'allègement, le domaine est étudié exhaustivement au-dessus d'un certain seuil, par sondage en dessous (cette procédure peut poser quelques problèmes en matière d'apurement de fichiers).

La troisième préoccupation est relative à l'amélioration de *la qualité* même des statistiques sur les entreprises. En effet, les erreurs que l'on peut commettre dans ce domaine proviennent pour l'essentiel de trois raisons :

- unités absentes (lorsque le fichier qui sert à lancer l'enquête est incomplet);
- unités mal classées (lorsqu'il y a erreur sur l'activité principale);
- réponses erronées.

Les deux premières causes sont de loin les plus importantes. Elles sont à l'origine de la plus grande partie des contradictions que l'on peut observer en comparant des informations d'origine différente.

La fusion des fichiers permet de corriger ces erreurs. En repérant les unités présentes dans un fichier et absentes dans l'autre, on découvre les unités absentes. Les classements peuvent être confrontés, et les cas douteux sont examinés avec plus de soin. Les divergences entre des réponses qui auraient dû être identiques permettent de repérer les erreurs. La fusion de fichiers apparaît donc comme *l'outil fondamental* qui permettra d'assurer solidement *la qualité* de la statistique des entreprises.

Elle pose des problèmes techniques très délicats, dont une bonne partie sera d'ailleurs rendue plus facile par l'utilisation de S. I. R. E. N. E. Pour le moment, l'expérience des S. U. S. E. est limitée à la fusion entre les B. I. C. (déclarations sur les bénéfices industriels et commerciaux, document administratif fourni par la Direction générale des impôts) et les enquêtes annuelles d'entreprises exécutées par les ministères techniques et L'I. N. S. E. E.

S. U. S. E. est appelé à servir de « noyau » à des banques de données spécialisées. Le projet le plus avancé (ENEIDE) est celui du ministère du Développement industriel et scientifique. Il fusionnera la partie de S. U. S. E. relative aux entreprises industrielles avec d'autres sources sur ces entreprises : enquêtes de branche, aides de l'État, etc. ⁽¹⁾.

c) Pour les unités géographiques

Il existe un premier fichier, il s'agit d'un fichier national topographique (FINATO) qui décrit l'ensemble des communes et rues. La référence de base est le segment de voie. Ce fichier est incomplet actuellement, car il ne concerne que les communes de 5 000 habitants et plus.

Le second qui constitue un projet est un fichier des coordonnées des parcelles cadastrales et des éléments descriptifs qui y sont attachés : Fichier topométrique (FITO). Il est envisagé également la constitution d'un système de localisation électronique (SILOE) dont le but serait de disposer pour chaque parcelle foncière de ses références cadastrales et de ses localisations les plus usuelles (adresse postale, coordonnées). Ce système se placerait à

1. Ministère du Développement industriel et scientifique. Analyse fonctionnelle de l'alimentation de la base des données, mars 1973.

l'articulation de fichiers particuliers relatifs à des caractéristiques foncières et rappellerait l'orientation suivie déjà en Suède et dans différents pays étrangers ⁽¹⁾.

Est actuellement en cours de constitution un répertoire géographique urbain (R. G. U.) qui décrira l'ensemble des îlots et même des faces d'îlots (segments de voies) et permettra ainsi de faire correspondre à une adresse postale un n° d'îlot. Ce répertoire sera cohérent avec le fichier FINATO.

3.3. *Les obstacles à cette approche*

Que peut-on penser de cette approche récente concernant les fichiers et les systèmes de données? Le rapprochement de données numériques issues de sources diverses est une des conséquences de la complexité de l'organisation administrative et de la multiplicité de ces sources ⁽²⁾. Il en résultera à l'avenir un effort non contestable d'homogénéisation des concepts, d'unification des nomenclatures et des identifiants des unités statistiques de base.

Néanmoins, il ne faut pas se dissimuler les grandes difficultés de cette approche. Outre le caractère fréquemment incomplet des fichiers administratifs, on peut noter :

— *Qualité du document :*

a) certaines réponses fournies aux services administratifs peuvent être orientées : exagération de certaines superficies au détriment d'autres, sous-estimation de cheptel ou de main-d'œuvre;

b) les rubriques non nécessaires aux opérations réglementaires peuvent être non remplies ou mal vérifiées par les agents de l'Administration. Une étude expérimentale a été réalisée par des chercheurs de l'Institut national de la recherche agronomique (station de Rennes) en collaboration avec le Service central des enquêtes et études statistiques (S. C. E. E. S.) du ministère de l'Agriculture, en vue de l'utilisation des statistiques fiscales pour déterminer l'importance des revenus non agricoles des agriculteurs bretons, années 1965-1966-1967. On s'est heurté à nombre de difficultés : non identité du ménage et du foyer fiscal, considération des seuls revenus imposables, non déclaration systématique de revenus, forfait des ressources discutables, pourcentage important de dossiers incomplets;

c) manque de stabilité du formulaire : sa forme et son contenu peuvent varier dans le temps et dans l'espace (documents de la Sécurité sociale par exemple);

d) les modifications de la réglementation peuvent entraîner la disparition de certains documents (déclaration en vue de l'obtention de la ristourne sur le matériel acheté par les exploitants agricoles);

e) la réglementation fiscale peut entraîner une fiabilité différente des documents : entreprises au forfait, entreprises au bénéfice réel.

— *non coïncidence* de l'unité statistique et du déclarant administratif : deux personnes déclarent au lieu et place d'une seule; le déclarant ne correspond pas au responsable économique réel. Ce phénomène est aggravé dans le domaine agricole par le fait qu'il n'existe pas une définition juridique de l'exploitation agricole.

1. Bjerke, An integrated social and demographic statistical system. Fortryck ur Statistisk tidskrift, 1970-3. Conference on small area statistics. American statistical Association, Pittsburgh, août 1968.

2. Dans le domaine des transports, plus de dix organismes divers se répartissent le rassemblement et l'exploitation de ces statistiques (Marine marchande, Aviation civile, Transports terrestres, Direction des routes, etc.). Dans celui de l'emploi, les sources sont également diverses : Sécurité sociale, Déclarations annuelles sur les salaires, Enquête trimestrielle du ministère du Travail, Enquête annuelle d'entreprises, etc.

Ceci n'est d'ailleurs pas propre aux seuls fichiers administratifs :

La comparaison de deux enquêtes statistiques peut être extrêmement révélatrice de l'incidence de la définition de l'unité statistique de base : un exemple très illustratif est fourni par la comparaison d'un recensement de la population et d'une enquête de structures en matière agricole. La donnée étudiée est celle du nombre d'exploitations agricoles ⁽¹⁾.

Au recensement de la population, on ne saisit que les chefs de ménage ayant déclaré comme activité principale l'activité agricole alors qu'à l'enquête agricole le balayage est plus exhaustif (une exploitation agricole peut être dirigée par une personne dont l'activité principale n'est pas agricole).

	Nombre total (1 000)
Au recensement de la population (1962)	1 772
A l'enquête agricole (1963)	1 899

Sur un échantillon représentatif de communes, on a été amené à préciser les origines de ces divergences ⁽²⁾.

Rubriques	Communes rurales	Communes urbaines	Ensemble
Exploitations communes aux deux opérations	85,7	60,5	84,7
Omissions au R. G. P. (en particulier le titulaire s'est déclaré retraité au recensement démographique)	7,2	10,6	7,4
Divergences dues à des déclarations non adéquates de la profession principale	7,1	28,9	7,9
	100	100	100

Nombre absolu d'exploitations confrontées : 18 000.

Remarque : l'opération de comparaison a permis de repérer 1,3 % d'omissions à l'enquête de structure.

1. Dans le même domaine, un exemple classique de divergence due à des concepts différents est celui du nombre d'exploitants agricoles enregistrés à la Mutualité sociale agricole (cotisants exonérés et non exonérés) et celui décompté par les services statistiques.

	Milliers d'exploitations		
	1955	1963	1971
Mutualité sociale agricole (A)	2 669	2 282	1 637
Statistique Agricole (B)	2 284	1 899	1 545
$\frac{A}{B} \times 100$	117	120	106

Exploitations dites « sans terre » non comprises.

Sur l'inadéquation pour certaines études économiques des renseignements détenus par la Mutualité Sociale Agricole on se reportera à « Utilisation des informations de la Mutualité sociale agricole en Économie rurale » par Ph. Lacombe, Institut national de la recherche agronomique, Montpellier, 1969.

2. Cahiers mensuels de statistique agricole, juillet-août 1966, S. C. E. E. S.

— Existence de seuils d'assujettissement différents : ce cas est extrêmement fréquent dans le domaine de la statistique agricole. Risque de non comparabilité chronologique en raison des variations de la législation.

— Incorporation progressive des données dans un fichier administratif en raison du décalage entre l'instruction du dossier et la décision finale. La mise à jour du fichier administratif est donc échelonnée, et l'amélioration épisodique ou localisée d'un fichier aggrave la difficulté de son utilisation (ce cas s'est rencontré fréquemment dans le passé, dans le cas de fichiers de vaccination du cheptel bovin).

Autrement dit, le rapprochement de sources administratives différentes peut être rendu malaisé du fait que la fusion des fiches peut faire ressortir les différences de numérotations des unités de base et des domaines d'études, des omissions, des différences de localisation géographique des unités statistiques ou même des différences de structure d'unités économiques de production. Les causes de ces non-appariements sont donc multiples :

- erreur d'identification numérique;
- différence de champ;
- différence de concept;
- différence de situation chronologique (les deux fichiers n'étant pas rigoureusement établis à la même date);
- différence de déclarations;

et il a paru utile de citer quelques expériences récentes en ce domaine :

En matière de *fichiers de personnes*, on a recherché pour le compte de deux services qui fournissaient le nom, les prénoms, le sexe, la date et le lieu de naissance de personnes, les identifiants correspondants. Les taux de réponses obtenus sont très élevés, cela tient à une programmation spécifique mais lourde. Chacune des informations élémentaires fait l'objet d'une certaine imprécision dans sa définition (notamment de lieu de naissance). Une comparaison brutale amènerait des résultats moins satisfaisants. L'un des buts de S. A. F. A. R. I. est donc de décharger les gestionnaires de fichiers de ces problèmes d'identification sur critères incomplets ou incertains. Il facilite l'accès aux fichiers et le rapprochement de fichiers de sources diverses. Pour le premier organisme, on a enregistré les résultats suivants :

a) demandes rejetées dès les premiers contrôles.	1,3 %
b) demandes identifiées.	94,6 %
c) demandes non identifiées.	4,1 %

Parmi les demandes identifiées, il subsistait des divergences importantes concernant les prénoms (90 % de ces divergences s'expliquaient par une liste incomplète des prénoms).

Dans le second cas, on trouvait (%) a) 6,5; b) 82,0; c) 11,5; là encore les divergences sur les identifiés (2/3 des cas) portaient sur une liste incomplète des prénoms.

En matière d'*entreprises*, on peut se rappeler un exemple emprunté aux enquêtes sur les industries agricoles et alimentaires.

Pour une activité particulière (charcuterie — salaison) on a été amené à rapprocher le nombre de salariés et le chiffre d'affaires, données provenant les unes d'une enquête annuelle dite secteur et orientées sur les données financières, les autres d'une enquête annuelle sur les productions physiques.

Pour l'année de référence (1969), on a les résultats suivants pour 381 entreprises qui ont pu être appariées sur 502 questionnaires étudiés.



Écart relatif pour le paramètre considéré ⁽¹⁾	Nombre d'entreprises identifiées	
	Salaries	Chiffres d'affaires
Identité	70	65
— 2 % d'écart	21	73
2 à 10 %	102	76
10 % et +	188	167
TOTAL	381	381

1. (résultats enquête sur les productions physiques — résultats enquête annuelle secteur)/
(résultats enquête sur les productions physiques).

Alors que l'enquête annuelle d'entreprise est exécutée par les services statistiques eux-mêmes, l'enquête sur les productions physiques est réalisée par les organisations professionnelles, lesquelles sont également chargées de percevoir une taxe parafiscale dont l'assiette repose sur le montant du chiffre d'affaires, il y a là une explication vraisemblable à la sous-estimation du chiffre d'affaires enregistré par le canal de l'enquête menée par les organisations professionnelles.

Une étude méthodologique réalisée par l'I. N. S. E. E. (Département des entreprises) a eu en 1970 pour objet de comparer deux sources exhaustives d'information ⁽¹⁾ : l'enquête annuelle d'entreprises (EAE) et les déclarations sur les bénéfices industriels et commerciaux, et ce sur les activités mines de fer, sidérurgie, automobile et travail mécanique du bois; on a constaté que 35 % possèdent des numéros identifiants différents (surtout parmi les grosses entreprises), 15 % des entreprises sont classées dans des secteurs d'activité différents. Au niveau des résultats publiés les divergences varient de 5 à 30 %; par contre, pour une même entreprise, les écarts pour une variable donnée n'excèdent pas en moyenne 2 %. Cet exemple illustre que l'un des problèmes de base est de classer une entreprise de façon identique quel que soit le canal de collecte des informations.

Dans le domaine des ménages, on a été amené à comparer à l'occasion de l'enquête logement 1967, pour un sous-échantillon commun les résultats bruts de cette enquête avec les éléments obtenus par dépouillement des dossiers des Caisses d'Allocations familiales des bénéficiaires de l'allocation logement. Si la comparaison globale des résultats est en général satisfaisante, la comparaison des paires d'informations au niveau individuel fait apparaître des écarts (même sur le nombre d'enfants à charge) par exemple : 12 % des locataires ont un loyer à l'enquête qui diffère de plus de 10 % du loyer déclaré auprès des Caisses (incidence des charges).

On peut citer un dernier exemple, emprunté encore au domaine des entreprises : une enquête est réalisée par le ministère du Travail et l'I. N. S. E. E. sur les établissements industriels et commerciaux, sur la structure des emplois (ESE), tandis que l'enquête annuelle sur les entreprises (EAE) est exécutée par le ministère du Développement industriel et scientifique sur des variables, surtout financières. D'une part, on a été amené à étudier la fusion des deux fichiers de base; d'autre part, on a étudié dans ces deux enquêtes leurs paramètres communs : activité principale et effectifs totaux (cadres, ouvriers, non salariés).

- a) On constate que pour les effectifs, il existe diverses sources de divergences :
— décalage temporel des enquêtes (3 mois);

1. Présentation des premiers résultats de la confrontation entre les B. I. C. et l'E. A. E. pour l'exercice 1968. Département des entreprises I. N. S. E. E.

- regroupements d'établissements;
- déclarations divergentes des entreprises, extrapolations dues aux sondages pour les unités de petite taille;
- corrections des non-répondants.

b) Si l'on examine le dénombrement des unités elles-mêmes, on constate qu'il peut y avoir divergences dans l'identification de l'établissement pour d'autres raisons :

- changement de numéro I. N. S. E. E. (activité principale) non ou mal enregistré dans chacun des deux fichiers;
- numéro de département différent (récent découpage de la région parisienne);
- implantation de l'établissement sur deux communes;
- erreur de transcription ou de perforation;
- établissements regroupés sous deux numéros différents;
- partie administrative et commerciale, différenciée dans l'une des sources de la partie usine.

Distribution relative de 22 000 entreprises classées selon la différence des effectifs salariés par établissement EAE-ESE (1969).

Ecart relatif sur les effectifs $\left(\frac{EAE - ESE}{EAE} \times 100 \right)$		Nombre relatif d'établissements (‰)
Différence négative supérieure à 50		38
— — comprise entre 50 et 20		44
— — comprise entre 20 et 10		70
— — comprise entre 10 et 5		85
— — comprise entre 5 et 0		169
Différence positive comprise entre 0 et 5		364
— — comprise entre 5 et 10		95
— — comprise entre 10 et 20		72
— — comprise entre 20 et 50		48
— — supérieur à 50		15
TOTAUX . . .		1 000

Si on considère le tronc commun des deux enquêtes après fusion des bandes, la comparaison des résultats pour chaque unité dénombrée doit tenir compte du phénomène sondage pour les petites tailles ⁽¹⁾ néanmoins, en considérant les établissements dont l'effectif est égal ou supérieur à 50 salariés, le recouvrement reste loin d'être complet :

	Nombre d'établissements	Effectif de salariés (1 000)
Nombre d'unités communes identifiées (UC)	10 387	2 690
Nombre d'unités présentes dans l'EAE seule	1 873	511
Nombre d'unités présentes dans l'ESE seule	2 154	482
UC/EAE	0,85	0,84
UC/ESE	0,82	0,85

L'apurement correct de fichiers constitue le préalable de base à des opérations de ce genre.

Cet apurement doit être ensuite complété par des procédures de mise à jour rapides et rigoureuses.

1. On peut, en effet, rencontrer la difficulté supplémentaire d'avoir à interclasser des données résultant d'une opération exhaustive et des données d'une autre source exhaustive seulement à partir d'un certain seuil (issues d'un échantillon au-dessous de ce seuil, il est alors nécessaire de raisonner par classe de taille).

D — AMÉLIORATION DE LA QUALITÉ DES DONNÉES

1. Il peut paraître surprenant de rappeler ce problème que l'on pourrait considérer comme résolu. En fait, il n'en est rien, et de loin. Les raisons en sont fort diverses; transmission partielle de l'expérience acquise, ambition de plus en plus grande des questions posées (sous la pression des utilisateurs), dégradation des réseaux d'enquêteurs (vieillesse des personnes, manque fréquent de personnel intermédiaire de contrôle), réticence des enquêtés (l'intensité de cette réticence est souvent « proportionnelle » à l'augmentation du niveau de vie de la classe sociale interrogée) ⁽¹⁾, la réticence des populations marginales est également très forte, modification des habitudes de vie des ménages (achats groupés, déplacements, indépendance des jeunes entraînant une individualisation d'achats mal enregistrés sur ressources propres). Dans le domaine des statistiques industrielles, la détérioration de la qualité de l'information est liée à la modification de la gestion : un groupe d'entreprises peut spécialiser ses filiales (unité chargée de la gestion du personnel, unités techniques, unité de gestion financière), la ventilation par secteur devient alors malaisée.

La liste de ces raisons n'est pas limitative, car on pourrait signaler aussi la difficulté de la réflexion méthodologique de base, faute de moyens et de temps.

2. On peut d'ailleurs se poser une question de fond qui est d'évaluer *le coût supplémentaire* envisageable pour l'obtention d'un gain donné de la qualité d'une information numérique déterminée. A l'usage, on s'aperçoit de la difficulté de quantifier cet arbitrage :

— les « avantages » d'une amélioration peuvent ne se faire sentir qu'avec un certain décalage de temps : la dérive d'un indice ne se perçoit pas immédiatement, l'utilité d'une information peut, selon la conjoncture, être notable à un certain moment, puis disparaître;

— dans le domaine agricole, le marché d'un produit peut être « cassé » par un excédent de la production sur les utilisations de 2 %, pour un autre produit (plus stockable), un excédent relatif plus important peut être toléré, et en conséquence, l'erreur statistique admissible peut être plus grande. Il est également fréquent que ce n'est pas tant le niveau absolu d'un estimateur qui soit recherché que la variation relative entre deux observations et il peut être moins coûteux d'améliorer la précision de cette variation;

— les « retombées » indirectes mais utiles d'une information peuvent ne pas être appréciées immédiatement.

En fait, l'expérience de tous les jours rappelle que ce sont les erreurs non aléatoires qui obèrent la qualité de l'information; si l'on expose en tête de ce chapitre quelques études ou données récentes en matière de sondages, il ne faut pas être dupe et être conscient qu'un plan de sondage s'écartant un peu de l'orthodoxie, mais rassemblant des données numériques

1. Enquête budgets familiaux 1963-1964. Refus selon la catégorie socio-professionnelle.

	Agriculteur exploitant	Ouvrier	Cadre sup. prof. libérale	Ensemble
	en %			
— ayant accepté de répondre à toute l'enquête (interview et carnet de compte)	81,7	67,9	59,1	67,2
— à l'interview seulement	10,5	15	17,8	15,0
— n'ayant pas accepté (refus au sens strict)	6,8	12,6	23,1	12,4

solides, suivies d'un traitement correct est souvent préférable à un échantillonnage très sophistiqué couvrant des relevés ou déclarations douteux, et suivi d'une élaboration ultérieure peu soignée du matériel numérique rassemblé. Il doit y avoir un arbitrage adéquat, lors de la réflexion, de l'affectation des moyens entre ces deux pôles également importants.

3. *Précision des estimateurs*

En matière de calculs d'erreurs liés au divers plans de sondage adoptés, on rappellera les deux études parues dans *ANNALES de l'I. N. S. E. E.* d'octobre 1969 — Fernand CHARTIER : « Erreurs d'échantillonnage pour divers plans de sondage de logements » — Jean-Louis BODIN : « Une comparaison expérimentale de la précision obtenue par divers tirages systématiques de groupes ». La première traite de la précision des estimations relatives à trois plans de sondage (un seul degré, plusieurs degrés, échantillons d'aires, la seconde traite des résultats obtenus en simulant des tirages d'échantillons sur ordinateurs à partir du recensement de la population de 1962 (tirages d'individus, de logements, d'immeubles).

Dans le domaine de la statistique agricole, on peut citer les travaux récents suivants :

— Travaux méthodologiques en vue du remplacement par trois enquêtes durant la campagne, de l'enquête annuelle sur le cheptel porcin. On a étudié en particulier l'évolution de l'erreur aléatoire selon la taille de l'échantillon. Le calcul a été régionalisé.

— Calcul d'erreurs sur les enquêtes :

- production fruitière 1969-1970;
- utilisation du territoire 1971 (analyse de la variance interphotographies aériennes et intraphotographies par culture et groupe de cultures);
- étude communautaire 1967-1968 (analyse de la variance par culture et par région);
- cheptel bovin au 1^{er} janvier 1972 (précision attachée à l'estimation du total du cheptel selon la taille de l'échantillon).

La régionalisation des enquêtes entraîne un alourdissement considérable des échantillons. Pour prendre un exemple concret, l'enquête sur la structure du cheptel porcin de décembre 1971 déjà citée, portait sur 19 000 exploitations, pour la rubrique « total porcs » le double du coefficient de variation s'élevait à 1,8 %, tandis que le même estimateur au niveau d'une région (21 au total) avait pour valeur minimum 4,6 % et pouvait dans certaines régions à spéculation peu porcine atteindre et dépasser 10 %. Dans le domaine des statistiques industrielles, l'obtention de données régionalisables entraîne une majoration de coût de l'ordre de 25 à 30 %.

4. *Les enquêtes par correspondance*

On a déjà vu dans la partie réservée aux renseignements fournis par les entreprises, certaines lacunes de l'information. Les non-réponses, les réponses partielles, les réponses biaisées, les retards des réponses constituent autant d'écueils. L'enquête par correspondance qui s'adresse à des ménages (par opposition à l'entretien par l'enquêteur) a déjà fait dans le passé l'objet de critiques dont celle de conduire fréquemment à des réponses sous-estimées.

Il faut cependant remarquer que dans certains cas, notamment lorsque l'enquête est spécifique et limitée à une recherche très particulière, l'enquête peut modifier le comportement du ménage et entraîner au contraire une sur estimation des résultats. Un exemple peut en être donné pour l'enquête « Fruits et Légumes de juillet-septembre 1969 » faite par l'I. N. S. E. E. (Division enquêtes de consommation et comptes des ménages) auprès de ménages en vacances au cours de laquelle on s'est aperçu de cet « effet de polarisation ». Comparée à une

enquête permanente alimentaire classique (enquêtes et carnets de ménage), l'enquête pilote par correspondance, visant des personnes non parties en vacances, a entraîné une sur estimation de la dépense de consommation de fruits et légumes selon les régions de 64 à 140 %, la moyenne nationale étant de 106 %. Les différences de structure des deux échantillons ont été analysées et ne constituent pas la source de cet écart significatif qui résulte surtout du comportement des ménages enquêtés par correspondance prenant soudain conscience du non-respect d'une « norme de consommation » tacite ⁽¹⁾.

5. Repérage des erreurs et contrôle sur le terrain

Les procédures de contrôle des enquêtes sur le terrain sont malaisées à mettre en œuvre et, pour chaque cas, il faut y consacrer une réflexion attentive :

- Accompagnement de l'enquêteur (par un contrôleur).
- Répétition de l'observation ou de l'entretien sur un sous-échantillon :
 - pour l'ensemble de l'enquête;
 - ou pour certaines rubriques « sensibles » du formulaire.
- Envoi par correspondance de cartes aux enquêtés leur demandant confirmation du passage de l'enquêteur et réponse à une question annexe : il arrive trop souvent que les enquêtés ne répondent pas ou confondent plusieurs enquêtes d'organismes différents. La fiabilité de cette procédure est très douteuse.
- Observation d'un échantillon indépendant par un réseau d'enquêteurs plus qualifiés.
- Examen pour certaines rubriques communes des « enquêtés-inter-sections » de deux enquêtes indépendantes. Cette technique a été appliquée en agriculture pour rapprocher des déclarations faites au R. G. A. de celles enregistrées dans les enquêtes céréales et vergers réalisées au cours de la même campagne agricole.
- Vérification d'exhaustivité par procédures objectives d'échantillonnages aréolaires. Cette approche a été mise en œuvre pour le recensement démographique en 1962 dans 400 aires de 50 logements environ ⁽²⁾.
- Utilisation de formulaires différents entraînant une technique de remplissage particulière : ce procédé a été appliqué dans l'enquête de décembre 1968 sur la structure du cheptel porcin ⁽³⁾, au cours de laquelle on imposait pour un sous-échantillon d'exploitants

1. On rappellera ici seulement pour mémoire le phénomène de l'inflation apparente des dépenses durant les 3 ou 4 premiers jours d'enquête lorsque les ménages tiennent un carnet de comptes (voir *Revue de Statistique appliquée*, 1968, volume XVI, n° 4 : Exemples d'une analyse méthodologique des résultats d'une enquête auprès des ménages par J. Desabie et J. M. Rempp à propos de l'enquête 1963-1964 déjà citée sur les dépenses des ménages). Ce phénomène est provoqué par diverses causes : lassitude croissante de la ménagère et augmentation des omissions pour les produits moins fréquents, achats de produits stockables surestimés, aide de l'enquêteur le premier jour, etc.

Article	Rang du jour de relevé					
	Indice du nombre des relevés					
	1	2	3	4	7	14
Lait	108,4	100,9	102,1	99,1	96,8	96,4
Huile	133,5	126,1	113,8	104,1	91,1	95,3
Pharmacie	111,6	112,4	111,1	107,0	94,3	97,9
Blanchisserie	96,9	101,7	102,0	106,5	100,8	113,0

2. G. VANGREVELINGHE, *Études Statistiques*, n° 4, octobre-décembre 1963. Étude exhaustive du dénombrement.

3. Enquête de décembre 1968. Supplément série *Études*, n° 58, novembre 1969, ministère de l'Agriculture,

le comptage des pores dans les porcheries. Par cette méthode les enquêteurs ont trouvé un effectif total de pores supérieur de 3,2 % à l'effectif trouvé à partir de la simple déclaration des exploitants.

— Recours à des fichiers administratifs faisant foi (état civil). Cette approche peut être utile pour relever des erreurs d'enquêtes (soit de collecte, soit d'exploitation). Une étude publiée en 1968, à l'occasion du Symposium international sur les problèmes de reproduction de la population de Varna (25-30 septembre) par S. HEMERY et M. FEBVAY, avait pour objet à partir du recensement de la population de 1962, d'isoler les personnes nées en 1872 et avant. A partir des renseignements relevés sur les bulletins individuels du recensement, on tentait de se rapprocher de la mairie de naissance pour s'assurer que les déclarations correspondaient aux mentions des registres d'état civil. Les erreurs se divisent en trois catégories :

- Erreurs de déclaration 4
- Doubles comptes. 1,6 (en pourcentages)
- Erreurs de chiffrage et de perforation. 3,5

Au total sur 100 personnes apparaissant au recensement comme âgées de 90 ans et plus, 9 figurent à tort dans cette catégorie.

— Le contrôle des données peut également être réalisé par l'articulation de deux opérations statistiques, et ce, pour les unités statistiques qui leur sont communes : un exemple en est fourni dans la R. G. A., 1970-1971, au cours duquel une opération de contrôle a été exécutée au printemps 1971 à partir d'un réseau de 60 000 points implantés sur un échantillon de photographies aériennes.

Les enquêteurs sur le terrain repéraient la culture portée et remontaient à l'exploitant responsable de la parcelle. Il était alors possible de contrôler objectivement la valeur du fichier d'exploitations agricoles et les déclarations des exploitants.

Les résultats de ces contrôles au niveau France entière ne sont pas disponibles, néanmoins quelques dépouillements départementaux confirment la richesse de l'information supplémentaire obtenue :

— Omission d'exploitations :

- généralement le chef n'a pas l'activité agricole comme activité principale;
- exploitations en bordures de limites administratives.

— Divergences diverses :

- au niveau du nom de l'exploitant;
- au niveau de la commune siège de l'exploitation;
- en ce qui concerne la déclaration de superficie agricole utile totale;
- absence d'une rubrique;
- points tombant sur une exploitation vacante;
- au niveau de la gestion du recensement (oublis de perforation, de mise sur bande, anomalies de numérotation).

— En matière de contrôle de qualité ou plutôt de cohérence, *l'apport de l'informatique* n'est pas négligeable : établissement de catalogues d'incompatibilités et de chaînes de programmes, remplacements par redressements automatiques de données manquantes ou erronées (hot-deck). Par ailleurs, l'intervention de l'informatique peut dans certains cas soulager la tâche du personnel chargé des travaux manuels en évitant des contrôles préliminaires manuels, et en les remplaçant par l'intervention de ces agents au stade de messages d'erreurs annoncés par la machine. Inversement il faut néanmoins être conscient de la répugnance des agents à transmettre des questionnaires contrôlés seulement pour certaines

rubriques (identifiants du document de base par exemple) et contraints d'attendre ensuite les messages d'erreurs avec le risque de tous les jugements de qualité attachés au nombre et à la nature de ces messages. Une formation particulière du personnel doit être prévue : contrôle intelligent manuel de certaines rubriques au lieu d'un contrôle systématique global, correction de listes et établissement de bordereaux à la place de corrections sur documents, etc.

— Enfin la mécanisation des opérations entraîne la négligence des contrôles manuels, repousse dans le temps les contrôles et les corrections, d'où des délais plus longs et une incertitude sur la valeur des rectifications des messages d'erreurs. De plus, l'introduction de l'informatique dans les entreprises enquêtées entraîne parfois des retards et nécessite que le questionnaire soit envoyé suffisamment à l'avance, de telle sorte que l'informatique de l'entreprise s'adapte aux questions posées.

Il est également très pédagogique d'établir des comptages d'erreurs par rubrique ; ces comptages sensibilisent le personnel chargé des relevés et du contrôle en bureau et permettent au personnel de conception le repérage des rubriques les plus délicates. On constate généralement que les incompatibilités relèvent de deux catégories très différentes : la première regroupe les incohérences formelles, le non respect des conventions d'écriture et les omissions, tandis que la seconde comprend plutôt les erreurs de compréhension ou d'interprétation des enquêteurs et les erreurs de saisie sur cartes ou bandes.

6. Les refus

Le problème des refus est majeur : à partir d'un certain taux de refus, on peut craindre que le lot des répondants ne soit plus représentatif de l'ensemble original (même si on prend la précaution d'un redressement a posteriori ⁽¹⁾) ; par ailleurs, le réseau d'enquêteurs peut « prendre l'habitude » de cet état de fait et s'y résigner ; enfin, les enquêtés qui refusent de répondre peuvent faire tache d'huile et détériorer le climat. Bien entendu, on suppose que le volume du questionnaire reste raisonnable et ne constitue pas en lui-même un motif de rejet.

A cet égard, il faut faire une différence assez nette entre les enquêtes récentes du service statistique du ministère de l'Agriculture et celles de l'I. N. S. E. E. Pour les premières, grâce surtout à une implantation rapprochée (départementale) d'enquêteurs, un encadrement relativement dense de contrôleurs, des horaires plus réguliers qu'en ville, et de moindres déplacements, les taux de refus sont inférieurs à 1 % ; pour les secondes, au contraire, le réseau d'enquêteurs est de moindre qualité, le contrôle plus lointain, les questionnaires parfois plus délicats et les taux de refus sont plus importants : pour deux enquêtes récentes, on a noté les taux suivants :

— enquête habillement (1971-1972)	7,8 %
— enquête intention d'achats (janvier 1973)	7,5 %

Mais si l'enquête s'accompagne, en plus, de la demande de transcription d'un document laissé aux mains du ménage, le taux de refus s'élève rapidement.

	Enquête seule	Enquête + dépôt d'un carnet
— Enquête consommation alimentaire . . .	8,4 %	23,9 %
— Enquête conditions de vie des ménages . .	8,9 %	20,8 %

1. On rappellera les difficultés de démarrage du réseau d'information comptable agricole pour lequel il a fallu visiter plus de 12 000 exploitants pour en trouver finalement 3 000 qui acceptent de faire partie de cet échantillon non aléatoire.

7. Les redressements à la suite des non-déclarations

Dans le domaine agricole, on s'est attaché avec beaucoup de persévérance à éviter les rubriques non déclarées ⁽¹⁾. Pour ce faire :

— l'enquêteur n'hésite pas à poursuivre l'entretien avec l'enquêté de manière à obtenir une réponse vraisemblable (on risque alors de voir l'enquêteur se substituer à l'enquêté);

— ou bien lors du contrôle manuel (départemental, régional ou national) un spécialiste complète les rubriques manquantes après examen du questionnaire;

— ou bien encore et notamment dans les enquêtes par correspondance auprès des entreprises du secteur des Industries agricoles et alimentaires, une correspondance navette aux chefs d'entreprises réclame les informations omises ou douteuses.

Dans les autres domaines l'I. N. S. E. E. utilise fréquemment, au niveau de l'unité statistique observée, la méthode du redressement automatique (hot-deck).

Cette méthode tient compte de la connaissance de la répartition de critères communs en forte corrélation avec le critère faisant l'objet de non-déclaration et par le jeu de programmes d'instruction permet à l'ordinateur de remplacer les mentions non déclarées par des valeurs numériques.

Cette méthode a ses limites :

— procédure qui alourdit la mémoire de la machine;

— faux sentiment de confort et risque de difficultés d'interprétation de certaines rubriques à l'intérieur d'une petite zone géographique;

— influence de la nature de la séquence liée au tri : la variable pour laquelle il existe des informations non déclarées ne doit pas tenir lieu de critère de tri;

— risque de distorsion des résultats en cas de prise en considération de rubriques rares dans un échantillon à taux de sondage élevé : on risque soit d'atténuer le phénomène singulier, soit au contraire, de l'amplifier par une pondération artificielle.

8. La qualité du chiffrage

La qualité du chiffrage de certaines opérations de masse a fait l'objet d'une attention particulière. On peut citer l'exemple suivant extrait d'une note de travail de mars 1972 non publiée de M. J. C. LABAT. A partir du recensement de la population 1968, on avait constitué par recopie de bulletins individuels, un échantillon dit « panel démographique » se composant de toutes les personnes nées les 1-2-3-4 octobre quelle que soit l'année de naissance, panel destiné à des études longitudinales d'ordre démographique. Pour ce panel, on a comparé le résultat d'un chiffrage et d'une perforation totalement indépendants (sur deux sous-échantillons 1/20 et 1/5) par rapport aux cartes obtenues à partir des bulletins originaux.

La confrontation des deux fichiers a été opérée sur ordinateur et on s'aperçoit que dans le cas du sous-échantillon au 1/20 et pour l'ensemble de la France, 37 % des couples de cartes qui ont pu être appariées présentent au moins une différence.

1. Lorsqu'il s'agit de refus totaux, on est amené, soit à doubler le questionnaire aux caractéristiques les plus voisines (taille en superficie, âge, spécialisation par exemple), soit à opérer un redressement après stratification *a posteriori* de l'échantillon effectivement réalisé.

Toutes les rubriques ne sont d'ailleurs pas concernées identiquement :

Différence sur	Nombre moyen de divergences pour 100 couples de cartes appariées	
	1/20	1/5
État matrimonial	0,5	0,6
Diplôme enseignement général	1,4	1,2
Type d'activité	2,4	1,9
Professions	9,9	9,5
Activité économique	7,9	7,1

Des études sont en cours pour expliciter chaque divergence :

- erreur de recopie lors de l'établissement du bulletin du panel;
- erreur de chiffrement ou de perforation du R. P.;
- erreur de chiffrement ou de perforation du panel.

Une étude méthodologique du même genre a été réalisée sur le chiffrement de l'enquête « formation qualification professionnelle » de 1970, on a examiné sur un échantillon de questionnaires :

a) le chiffrement de la profession (à 4 chiffres);

b) le chiffrement de l'activité économique de l'établissement employeur (à 4 chiffres); et ce, à l'occasion de 5 rubriques pour a) (professions au Recensement de la population de 1968, au moment de l'enquête, du père de l'enquêté, etc.) et de 4 rubriques pour b).

Les écarts relatifs entre le chiffrement du personnel et le chiffrement de contrôle varient de 6,4 à 13,9 % pour a) et de 11,4 à 16,6 % pour b) (1).

9. Chiffrement automatique de certaines données

Dans toute la chaîne de travaux manuels qui sépare la collecte de l'information jusqu'à sa mise sur bande, intervient généralement le stade du chiffrement par des ateliers. Dans certains cas, les concepts et les codes sont simples et l'opération est relativement rapide (2). Dans d'autres, par contre, le personnel doit se livrer à une recherche parfois longue et délicate : l'exemple le plus illustratif est fourni par le chiffrement de l'activité économique de l'entreprise employeuse à l'occasion d'un recensement de population. Le personnel de chiffrement doit disposer de listes d'établissements classés par rues ou de listes de rues fournissant l'adresse des sièges d'établissements. Ces opérations manuelles sont lourdes dans les grandes agglomérations urbaines, et on risque la non-consultation de ces listes par les agents chargés du chiffrement.

Actuellement est étudiée la possibilité de disposer de mémoires (disques), lesquelles consultées en temps réel à partir de terminaux permettraient aux opératrices de poser les données connues (nom de l'établissement et adresse) et de disposer par consultation optique de l'activité économique principale (3).

1. Informations extraites d'une note de la Division Emploi de l'I. N. S. E. E., février 1972.

2. Elle peut d'ailleurs être accélérée par une saisie sur bande instantanée et sans chiffrement intermédiaire (saisie « on line »).

3. Une procédure semblable peut être envisagée pour la codification automatique du lieu de travail,

Il faut regretter plus généralement que le problème du chiffrage n'ait pas fait l'objet de recherches méthodologiques systématiques : cette opération délicate exige une formation correcte du personnel et des procédures de contrôle adaptées à chaque type d'enquêtes.

La codification automatique peut être également employée pour définir une position de codification à partir d'informations composantes. On pourrait penser par exemple déterminer la catégorie socio-professionnelle ⁽¹⁾ d'une personne à partir de la profession (activité individuelle) et du statut (indépendant, employeur, aide familial, apprenti, salarié secteur privé, salarié secteur public). Les expériences menées jusqu'ici semblent entraîner, comparées à une codification directe, certaines erreurs systématiques (surestimation du nombre des ouvriers, sous-estimation des effectifs des salariés agricoles, des membres des professions libérales, des cadres moyens et d'employés) ⁽²⁾.

10. La modernisation de la saisie

Sous ce terme, on comprendra les procédures tendant à se passer de la carte perforée.

Lecture optique : ce procédé n'est pas nouveau, mais il est vraisemblable que, soit pour des documents simples ou légers, soit pour des opérations répétitives et stables ⁽³⁾ cette méthode peut être rentable. Signalons que la lecture optique a été utilisée :

a) pour une partie de l'enquête par sondage sur le cheptel porcin du 1^{er} décembre 1970 (voir supplément, série *Études*, n° 81, ministère de l'Agriculture);

b) pour l'enquête pilote réalisée en juin 1972, dans deux quartiers de Lyon, en prévision du recensement de la population de 1975 (30 000 personnes environ).

Cette méthode a l'avantage de raccourcir les délais, de diminuer l'appel à du personnel temporaire de chiffrage; par contre, à chaque application, il y a lieu de se poser le problème des coûts comparés et de la formation particulière des enquêteurs (écriture) ⁽⁴⁾ est nécessaire. Il est par ailleurs souhaitable, que lors de la mise sur bande, le lecteur soit couplé à un ordinateur pour un premier passage de tests.

11. Saisie directe sur bande ou encodage

Un nouveau matériel permet de reconvertir le personnel chargé de la perforation classique pour le former à partir de la lecture du document de base à constituer directement des bandes magnétiques supports des données originales. Ces bandes doivent être évidemment ensuite triées, puisque l'entrée se fait généralement « en vrac ». Comparé à la perforation

1. Le code des catégories socio-professionnelles est destiné à permettre le classement de la population active en un nombre restreint de grandes catégories présentant chacune une certaine homogénéité sociale.

2. Comparaison des résultats du 1/4 et de l'exhaustif du R. P. 68 en ce qui concerne les catégories socio-professionnelles. Statistique générale I. N. S. E. E., novembre 1972.

3. Saisie des opérations du contrôle de la qualité et de la répression des fraudes au ministère de l'Agriculture par exemple.

4. Une expérience comparative (enquête sur la structure du cheptel porcin) a montré que la presque totalité des erreurs semblent provenir de caractères écrits, non standard pour le lecteur optique, et notamment pour le caractère 9.

Chiffres	0	1	2	3	4	5	6	7	8	9
Erreurs (%)	0,5	3	2	1	4	4	5	1	2	20

classique, le rendement serait de l'ordre de 20 % supérieur. Ce matériel présente également l'avantage de pouvoir être doté d'un programme de contrôle qui permet au moment de l'imposition des données de s'assurer avant enregistrement définitif de certaines cohérences.

La généralisation de l'usage de ce matériel à l'I. N. S. E. E. devrait être rapide, et dès 1975, il ne resterait que quelques machines classiques (pour la perforation de programmes par exemple).

CONCLUSION

Toute organisation ne peut être figée, ni dans ses structures, ni dans ses finalités, aussi les six observations qui suivent n'ont pas la prétention d'être définitives.

Leur reprise en considération est à la base du développement, entendu au sens de la qualité des données :

— Le réseau d'enquêteurs constitue déjà à lui seul tout un problème de fond : formation, conservation, intérêt de ces agents sont les conditions préalables à l'existence de ce réseau.

— La structure des questionnaires doit être repensée : l'enquête pilote doit être de règle, comporter un nombre important de rubriques, faire l'objet d'une analyse des données afin de permettre la sélection des rubriques du questionnaire définitif; ces rubriques visant, soit à fournir des inventaires, soit à mettre l'accent sur les critères explicatifs (en vue d'études sur tableaux, données typologiques, corrélations, etc.).

— Les ateliers manuels de contrôle et de chiffrage doivent être repensés sur le plan de l'organisation, car à l'ancien système de découpage en tâches élémentaires, on doit essayer de substituer une chaîne de tâches reconstituées. Le personnel chargé du chiffrage doit être plus qualifié.

— La saisie directe des données est en pleine évolution et l'importance de la carte perforée devrait aller en diminuant.

— L'intervention de l'informatique, une fois les premiers enthousiasmes éteints doit être repensée :

- sélection plus sévère des tabulations et exploitations recherchées : le statisticien ne doit pas crouler sous le poids du papier;
- stabilisation de l'emploi des matériels : trop souvent le client est acculé à un changement de machine sous le prétexte d'améliorations technologiques; stabilité également des programmes généraux d'exploitation; la modification d'un programme est parfois génératrice de perte de temps et nécessite un réinvestissement en matière de formation.

La maîtrise de l'informatique sous tous ses aspects devrait alors faciliter le dialogue statisticien-informaticien et la réduction des délais d'obtention des résultats.

— Enfin, une dernière observation, plus qualitative celle-là, doit être faite.

L'époque n'est pas éloignée où l'exercice de l'enquête (exhaustive ou par sondage) constituait une part très importante de l'activité des Instituts ou Services de statistique, le statisticien, depuis, a été tenté par les études d'aval, « noble couronnement de son activité », tandis que parallèlement se développaient les recherches pour tirer de la gestion administrative les fichiers ou les sous-produits utiles à l'information socio-économique. Dans cet univers en expansion, on a un peu l'impression que la collecte sur le terrain des données

numériques ne bénéficie pas d'une haute priorité. Il faut souligner au contraire, d'une part, qu'il reste encore des domaines de recherche justiciables de cette approche et d'autre part, qu'un investissement méthodologique important doit être consenti afin d'assurer un éclairage correct aux voies qui seront choisies.

G. THÉODORE, *Inspecteur général à l'I. N. S. E. E.*
en collaboration avec J. M. REMPP
Administrateur à l'I. N. S. E. E.