LES PROBLEMES RELATIFS AU TRAITEMENT INFORMATIQUE DES DONNEES

****

DOMENACH Hervé
ORSTOM - Martinique

La présente communication s'appuie sur l'expérience acquise au cours de la réalisation d'enquêtes sur l'emploi dans les quatre départements d'outre-mer Français : Guyane (1976), Réunion (1978-79), Martinique (1979-80) et Guadeloupe (1980), ainsi que des confrontations méthodologiques et analytiques avec des études similaires, notamment celles menées dans le bassin Caraïbe (Trinidad et Tobago, Barbade, Porto Rico, Jamaïque...).

1 - LA PROBLEMATIQUE INITIALE :

I.1 - Les contraintes matérielles :

Ce sont malheureusement plus souvent elles qui conditionnent les choix méthodologiques que l'inverse. Les contraintes budgétaires sont évidemment déterminantes des choix d'analyse , puisqu'impliquant une sélection de la tabulation et donc des variables et de leurs modalités, sauf à limiter également la collecte des données sur le terrain. On prendra également en compte les contraintes de temps qui dépendent à la fois des données physiques de l'espace étudié pour la collecte, et les moyens matériels et humains mis au service du traitement des données.

I.2 - Les contraintes conceptuelles :

Se pose le problème de pouvoir rapprocher les résultats obtenus des divers autres indicateurs statistiques concernant les lieux étudiés, mais aussi des séries internationales, aux fins de comparaison dans l'espace et dans le temps. S'il s'agit d'une étude aux fins d'analyse conjoncturelle, avec référence à des travaux antérieurs similaires, la chaîne collecte-traitement-analyse est en principe largement maîtrisée a priori et n'est guère soumise à des modifications. Il en va différemment s'il s'agit d'une étude aux fins d'analyse structurelle ou encore consacrée à un sujet spécifique nouveau ; en effet, la recherche de corrélations nouvelles suppose des données nombreuses, des modalités relativement ouvertes et des possibilités de traitement approfondi : exploitations informatiques complémentaires en fonction des résultats obtenus, recherche systématique de corrélations...

II - LES SOLUTIONS POSSIBLES :

II.1 - Concernant le questionnaire :

Une solution intéressante consiste à le concevoir par sous-ensembles ou modules, qui sont déterminés par les diverses sous-populations que l'on veut retrouver à l'analyse. Les caractéristiques de l'exercice d'un emploi par exemple : la profession, le statut, l'activité économique de l'entreprise, la catégorie socio-professionnelle, le lieu d'exercice et la durée travaillée... seront appréhendés de manière identique pour l'emploi actuel, l'emploi précédent, le premier emploi exercé, l'emploi souhaité... On peut également chercher à ce que chacune des principales catégories d'activité permette une analyse indépendante, soit à dissocier au maximum les modules concernant l'activité, le sous-emploi, la demande de travail et l'inactivité.

On peut alors utiliser un système de feuilles séparées, mais le questionnement numéroté avec sauts et renvois paraît le plus fonctionnel, ménageant un tronc commun propre à chacun des thèmes abordés : la formation, la mobilité... Si les questions sont fermées, avec des modalités de réponses codées, il est intéressant de prévoir des regroupements en continu des modalités détaillées, de manière à avoir par la suite plusieurs niveaux de tabulation possibles de la variable étudiée.

II.2 - Concernant la codification-saisie :

L'utilisation d'un fichier codifié parfaitement conçu s'avère primordiale pour le traitement des données. A la description détaillée des modalités de chacune des variables, il faut ajouter une modalité supplémentaire, toujours la même (soit 00, soit de l'alpha...) correspondant à la situation "non concerné" ; ainsi, l'on sera toujours en mesure de pouvoir vérifier le complément à la population totale. Les regroupements dont on a parlé au paragraphe précédent, peuvent être faits soit sur le premier chiffre lorsque les modalités sont nombreuses, soit en continu ; cette deuxième solution paraît préférable dans la mesure où l'on diminue les instructions de programme.

L'usage des variables calculées (ou encore de variables synthétiques, c'est-à-dire construites à partir de plusieurs variables : type de ménage, catégorie d'activité...) est tout à fait recommandable , car grandement simplificateur lors de la tabulation et du traitement. On a même intérêt à utiliser cette formule pour certaines variables qui seront souvent tabulées avec les mêmes regroupements : cas de l'âge par exemple qui est fréquemment croisé par groupes quinquennaux ou duodecennaux.

La codification directement sur le questionnaire est largement avantageuse à de nombreux égards : facilités de contrôle ou redressement ultérieur, reprise directe des codes utilisés pour les réponses fermées, sauts systématiques selon les modules concernés...

La saisie-écran conversationnelle, en attendant la saisie simultanée lors de la collecte, donne pleinement satisfaction lorsqu'elle

s'appuie d'une part sur les contrôles de validité habituels, et d'autre part sur des contrôles de cohérence utilisant notamment les situations d'exclusion de champ, c'est-à-dire les sous-populations non concernées par telle ou telle corrélation.

II.3 - Concernant la tabulation et l'analyse :

On peut utiliser deux conditions préalables permettant d'apprécier la signification analytique des données :

- opérer un dénombrement pour chaque variable de toutes les modalités prévues dans le fichier codifié ; selon les effectifs obtenus, on peut ainsi être amené à ne travailler qu'avec des modalités regroupées, ou tout simplement à annuler l'exploitation de certaines variables dont les effectifs seraient jugés non significatifs,

- déterminer pour chacun des tableaux demandés un "nombre moyen d'individus par case", qui est le rapport entre la sous-population concernée et l'ensemble des situations possibles ; en d'autres termes, il s'agit de diviser le nombre total d'individus compris dans le champ du tableau programmé, par le nombre total de cases que l'on obtient en multipliant le nombre de lignes par le nombre de colonnes. On aura par ailleurs déterminé un seuil de signification de cet indice : xx individus par case, qui est fonction de la fraction de sondage, de l'importance relative du thème du tableau, etc. On pourra ainsi, soit éliminer tout simplement les tableaux dont l'indice serait inférieur à ce seuil, soit au minimum travailler par priorités successives dans le traitement des tableaux prévus.

Il faut également pouvoir se référer à un plan de tabulation synthétique, qui fasse apparaître tous les croisements envisagés. Pour ce faire, on peut dissocier les tableaux internes à un module, c'est-à-dire à chaque sous-population, des tableaux intermodules ou d'ensemble. On peut ainsi répertorier les différentes variables en colonnes, et porter en lignes l'identifiant de chacun des tableaux obtenus pour chaque croisement de variables pointées dans les cases correspondantes. Au total, on obtient ainsi un nuage de croix centré sur la diagonale principale dont on peut rapidement apprécier la régularité, les excroissances et les vides éventuels.

II.4 - Concernant l'exploitation informatique :

Si l'on passe par sous-traitance, on se heurte inévitablement à l'inertie des instructions préalablement formulées ; l'analyste-programmeur se plaçant alors souvent en position d'exécutant sans capacité critique, et programmant plusieurs tableaux ensemble aux fins de simplification de son travail, il n'est guère possible de réajuster la demande au fur et à mesure des résultats obtenus ou des erreurs relevées. Si cela n'a guère de conséquences dans le cas d'une étude conjoncturelle effectuée en série, il en va tout à fait différemment pour une étude de type structurel ou spécifique, qui fait une large part à la recherche.

Etre en mesure de traiter soi-même les données constitue un avantage indéniable et le dialogue permanent avec l'ordinateur permet

de simplifier grandement le travail d'exploitation ainsi que d'éviter des frais inutiles. Une solution intéressante consiste à pouvoir disposer d'un progiciel spécifique au travail effectué, de manière à ce que la demande de tableau soit limitée à la sélection des variables et des filtres nécessaires. Si l'on peut de surcroît utiliser l'alphabétique, on a alors tout avantage à s'orienter vers une impression directement utilisable en publication.

En conclusion, on retiendra au travers de ces brèves considérations la nécessité de pouvoir traduire les concepts analytiques en variables clairement nomenclaturées d'une part, et l'avantage qu'il y a à travailler par modules séparés d'autre part.

MAROC
MINISTERE DU PLAN
DIRECTION DE LA STATISTIQUE

FRANCE
INSEE
ORSTOM

ASSOCIATION INTERNATIONALE
DES STATISTICIENS D'ENQUETES

SEMINAIRE SUR LES STATISTIQUES DE L'EMPLOI ET DU SECTEUR NON STRUCTURE

Rabat, 10-17 Octobre 1984
rapport des sessions et communications
tome 2

MAROC
MINISTÈRE DU PLAN
DIRECTION DE LA STATISTIQUE

FRANCE
INSEE
ORSTOM

ASSOCIATION INTERNATIONALE DES
STATISTICIENS D'ENQUÊTES

SEMINAIRE
SUR LES STATISTIQUES
DE L'EMPLOI
ET DU SECTEUR NON STRUCTURE

RABAT, 10-17 OCTOBRE 1984
RAPPORT DES SESSIONS ET COMMUNICATIONS
TOME 2

PARIS - JUILLET 1985