PRESENTATION D'UN GESTIONNAIRE DE DONNEES NUMERIQUES HIERARCHISEES DESTINE AU DEPOUILLEMENT D'ENQUETES

Jacques Vaugelade & Marie Piron
(Démographie et Statistique)
(UR 702)

Centre ORSTOM de Ouagadougou
B.P. 182
Ouagadougou (Burkina Faso)

RESUME - Le dépouillement d'enquêtes est la suite d'opérations qui conduit du questionnaire au tableau de fréquence (appelé aussi tableau statistique, croisé ou de contingence). Cette suite d'opérations comprend des modules obligatoires : description des variables, description des contrôles, saisie et correction, calcul des nouvelles variables et tabulation.

Le questionnaire présenté ici offre la particularité de traiter des questionnaires hiérarchisés et de permettre une saisie interactive. Il peut être complété et interfacé avec des logiciels statistiques qui réalisent des opérations du type régression, analyse de données multidimensionnelles, etc..

Dans ce cas la structure hiérarchique doit être abandonnée au profit d'une structure rectangulaire qui entraîne la répartition des informations des niveaux supérieurs pour chaque unité des niveaux inférieurs.

1. OBJECTIFS RECHERCHES

. Concevoir un gestionnaire le moins limité possible dans le nombre d'enregistrements, des variables, des individus et dans les possibilités de composer la tabulation.

. Saisir les informations, aussi complexes soient-elles, dans l'organisation du questionnaire qui se présente parfois sous la forme de différents niveaux d'enquête. Il s'agit d'enregistrer les données sans redondance des unités supérieures et pour un nombre quelconque de niveaux (principe d'une structure hiérarchisée).

. Etre attentif aux détails qui facilitent ou permettent d'exécuter des opérations qui font parfois défaut sur les autres logiciels de même profil.

. Enfin, concevoir un logiciel pour micro-ordinateur afin de pouvoir saisir et traiter immédiatement l'information localement, suivant les exigences précitées, aspect non négligeable pour les études portant sur les pays en développement.

2. PRESENTATION DU GESTIONNAIRE

Ce gestionnaire, conçu par J.Vaugelade, est écrit en Basic pour micro-ordinateur compatible IBM. Il permet actuellement de traiter uniquement des données numériques entières (valeurs comprises entre -32 768 et +32 767).

Il se présente sous la forme de menus déroulants et comporte les modules essentiels à son fonctionnement, à savoir :

2.1. Description des unités, des variables et des contrôles

Cette phase préliminaire consiste à décrire la structuration de l'enquête (un questionnaire peut être composé d'un ensemble de sous-questionnaires définissant chacun des unités secondaires, tertiaires, etc.. L'unité principale est l'unité statistique d'enquête dont peuvent dépendre plusieurs unités de niveau 2, 3, etc.). Il s'agit de définir les types d'unité et leur niveau dans la hiérarchie, de constituer les informations nécessaires pour décrire les variables qui leur sont affectées (libellé, nom abrégé, valeurs possibles) et d'établir des contrôles croisés entre les variables, basés sur des tests logiques.

2.2. Saisie, mise à jour, contrôle des données

La saisie est effectuée de manière interactive avec contrôle et correction immédiate.

Elle se particularise par une structure hiérarchique qui offre par là même deux avantages : une saisie simple compte tenu de la complexité du questionnaire pouvant faire intervenir différents niveaux d'enquête et l'occupation du minimum de place dans les fichiers d'enregistrement.

Lors de la première saisie, il suffit d'appeler l'unité désirée en respectant l'ordre de la hiérarchie, les variables concernées s'affichent alors.

Avec le programme de mise à jour, il est possible de se déplacer dans le fichier afin d'effectuer des corrections sur les variables, de supprimer ou d'ajouter des variables ou des unités. Ceci permet, par conséquent, d'enregistrer plusieurs passages d'enquête en ajoutant soit des variables, soit des unités statistiques d'enquêtes.

Cette saisie est simultanément accompagnée d'un prétraitement permettant de contrôler la qualité de l'information en décelant les erreurs de codification ou de saisie. Un premier

contrôle nous assure que chaque réponse est possible car elle appartient à une catégorie de codes précisés (Cf. description des variables), le deuxième contrôle, dit croisé, permet de vérifier la pertinence de la valeur de la variable (Cf. description des contrôles).

2.3. Création des variables

A partir des variables existantes, il est souvent nécessaire de créer par synthèse d'autres variables plus adaptées aux objectifs de l'étude. Par cette fonction, il est possible de synthétiser les informations d'un niveau inférieur ou d'utiliser les informations des niveaux supérieurs (on peut de plus faire intervenir dans le calcul les unités précédentes appartenant au même type d'unité de la variable concernée). Il est également prévu de créer une variable pour une sous-population constituée à l'aide de filtres.

2.4. Tabulation

Ce module permet de composer des tableaux de contingence (ou de fréquence) à partir d'une ou de plusieurs variables. Il permet de recoder ou de reclasser une variable, de lui affecter un intitulé, d'en sélectionner ses modalités, de ventiler un nombre quelconque de variables avec le choix de faire intervenir le total ou le rebut, de pondérer la population et enfin de travailler sur un type de population précis, sélectionné par des filtres.

Cependant, les types d'unités de même niveau sont indépendants entre eux. Par conséquent, les informations d'une unité peuvent être corrélées soit entre elles, soit avec les informations des unités supérieures dont elles sont dépendantes. Mais rappelons que nous sommes toujours à même de synthétiser l'information de deux types d'unités de même niveau au type d'unité supérieure commune.

D'autres modules existent comme lister ou trier des données suivant des variables prédéfinies, récupérer des fichiers endommagés et accéder à des fichiers séquentiels. En effet, concernant ce dernier point, il est à préciser que la structure hiérarchique des fichiers impose un enregistrement à accès direct.

La sortie séquentielle en caractères peut s'effectuer sous la forme de tableau rectangulaire ou hiérarchique. Ce module a toute son importance puisqu'il permet de passer sur des logiciels statistiques spécifiques. A cet effet, un passage vers STATITCF est actuellement en cours de réalisation.

3. EVALUATION

3.1. Propositions d'amélioration

. Etendre au traitement de données non entières et non numériques.

. Elargir les contrôles croisés, lors de la saisie, à des contrôles portant sur plusieurs variables, calqués en plus des tests logiques, sur le principe de création des variables.

. Perfectionner l'ergonomie et notamment au niveau de l'homogénéité dans les différents programmes.

. Optimiser la vitesse d'exécution des différents modules et du stockage des données.

. Réaliser une version compilée.

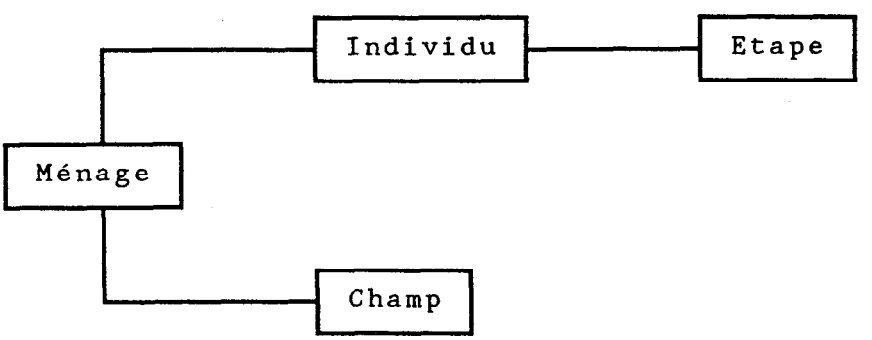
. Améliorer la présentation des tableaux croisés.

. Et pourquoi pas, étudier les possibilités d'une manipulation directe sur le tableau des données par le déplacement du curseur.

3.2. Points forts

L'originalité de ce gestionnaire est de pouvoir traiter des fichiers hiérarchisés à un nombre quelconque de niveaux. Lorsque les fichiers sont limités à deux niveaux, il est toujours possible de les ramener à un niveau unique en prévoyant le nombre maximal d'occurrences du deuxième niveau. Cela n'est plus possible au delà de deux niveaux. Nous attirons donc l'attention sur le principe d'une saisie hiérarchique qui consiste à dégager du questionnaire de base plusieurs sous-questionnaires correspondant chacun à un type d'information recueillie et pouvant, par conséquent, être considérés comme des sous-unités statistiques.

Si nous envisageons un fichier avec quatre types d'unités (Ménage, Individu, Etape, Champ) avec la structure suivante :



Chaque unité est caractérisée par un certain nombre de variables et dépend impérativement d'une unité supérieure qui l'identifie. Les unités secondaires sont enregistrées autant de fois

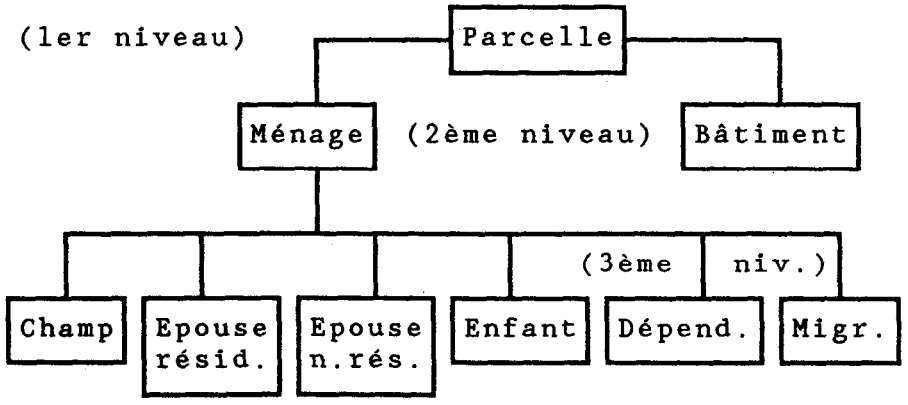
qu'il est nécessaire pour le questionnaire, en respectant l'ordre de la hiérarchie. Cette démarche permet, par conséquent, de focaliser progressivement les informations caractéristiques du ménage vers l'unité étape. Cette méthode de saisie présente l'avantage de traiter une catégorie d'enquête qui sera plus riche que les enquêtes restreintes à un seul type d'unité qui synthétisent et donc perdent l'information des sous-unités.

Ce gestionnaire offre un dimensionnement des fichiers en fonction des besoins de l'utilisateur, la seule limite relative reste la mémoire disponible de l'ordinateur, à savoir 64 Koctets pour le programme et les données.

Une grande souplesse d'utilisation est apportée par de nombreux outils de manipulation qui facilitent le traitement des données dont on augmente les possibilités (étude sur une sous-population sans revenir au tableau des données, existence de rebut, maniement des variables et de leurs modalités, etc.).

4. APPLICATION

M.Piron a utilisé et testé ce gestionnaire de données numériques pour une enquête socio-urbaine à passages répétés, comprenant neuf types d'unités réparties sur trois niveaux :



Elle n'a été contrainte à aucun dimensionnement, bien que le nombre de variables soit relativement important, plus de 300 variables et environ 13 000 unités confondues, après le deuxième passage de l'enquête. Elle a pu supprimer, modifier les variables, réaliser un premier traitement de données avant de passer sur des logiciels spécifiques tels que l'analyse multidimensionnelle, pour une étude plus approfondie des phénomènes urbains à Ouagadougou.

ANNEXE

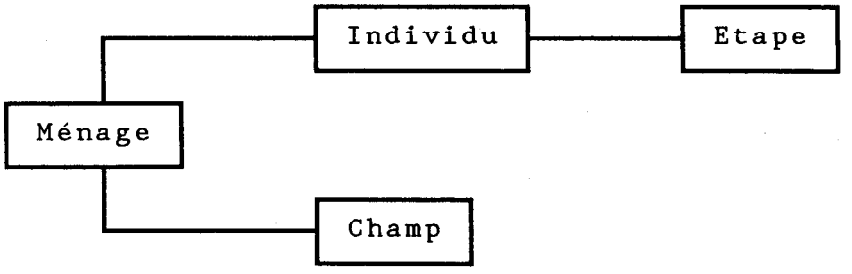
Structure des fichiers hiérarchisés : principe d'enregistrement

. Pour permettre la correction des unités et la mise à jour, le fichier doit être à accès direct, il faut alors choisir une longueur d'enregistrement. Plusieurs possibilités s'offrent, choisir l'unité ayant le plus de variables pour fixer la longueur de l'enregistrement conduit à une perte de place importante pour un micro-ordinateur ; créer un fichier par type d'unité présente deux inconvénients, le premier est une limitation du nombre d'unités par le nombre de fichiers pouvant être ouverts simultanément, le deuxième est une perte de temps en écriture et en lecture, car le temps de positionnement pour accéder à un enregistrement (secteur) est de l'ordre de cent fois le temps de lecture d'un secteur.

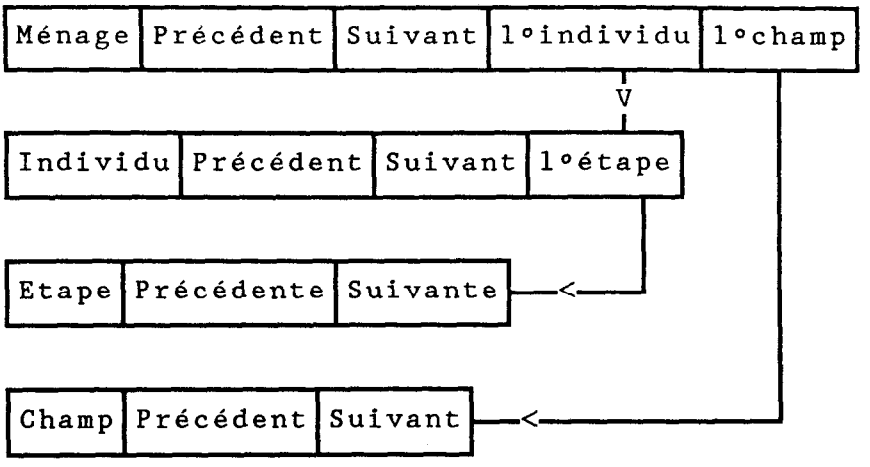
La structure retenue est donc une structure séquentielle, chaque enregistrement à une longueur fixe (par exemple 126 variables entières sur deux octets) et les unités sont mises les unes à la suite des autres.

. Pour permettre la mise à jour, il faut pouvoir se déplacer dans le fichier dans n'importe quel sens, vers l'avant ou vers l'arrière. On est donc conduit à introduire des pointeurs que l'on va illustrer par un exemple.

Soit un fichier avec quatre type d'unités, ménage, individu, étape, champ, avec la structure suivante :



Pour chaque unité on a un pointeur sur l'unité précédente et sur l'unité suivante, et pour chaque unité non terminale autant de pointeurs qu'il y a d'unités directement dépendantes.



Au cas où un pointeur est sans objet, sa valeur est zéro. Chaque pointeur donne le déplacement relatif, sauf si l'unité correspondante est dans le fichier des ajouts, auquel cas c'est le numéro de l'enregistrement (précédé d'un signe négatif) qui est le pointeur.