AFRISTAT
OBSERVATOIRE ECONOMIQUE ET STATISTIQUE DôAFRIQUE SUBSAHARIENNE

DIAL
DEVELOPPEMENT ET INSERTION INTERNATIONALE

PROJET PARSTAT

INSTRUCTIONS SUR LôAPUREMENT DES DONNEES DE LA PHASE 1

Septembre 2001

Introduction

Dans une opération telle que l'enquête 1-2-3, le traitement informatique se déroule en deux grandes étapes : la saisie et l'apurement des données. Après la phase d'apurement, la suite du traitement relève de l'analyse. Les instructions qui suivent visent à permettre l'harmonisation des procédures de traitement.

La règle générale est de ne pas attendre la fin de la collecte pour démarrer les opérations de saisie et d'apurement ; en fait la stratégie adoptée consiste à réaliser l'apurement pendant le déroulement des opérations de collecte. Si l'apurement ne commence qu'après que la collecte ne soit complètement achevée, on perd la possibilité d'effectuer des retours de terrain. ***Un bon timing des opérations serait d'engager la formation des agents de saisie une semaine avant le début de la collecte ou tout au moins au moment où débute cette phase de l'enquête.*** La saisie proprement dite commencerait alors une semaine après la collecte et l'apurement commencerait une semaine après la saisie ; c'est à dire deux semaines après le début de la collecte.

On propose que l'apurement de la phase 1 se déroule par vagues de 500 ménages qui seront dénommés LOT01 à LOT05 conformément aux instructions du manuel de l'agent de saisie. En fait, étant donné qu'on utilise 10 agents de saisie, les 10 fichiers (un par agent) du premier batch peuvent être dénommés LOT01A à LOT01J, les 10 fichiers du deuxième batch LOT02A à LOT02J, etc.

Pratiquement, après la saisie des 500 premiers ménages, les dix fichiers précédents d'un même batch sont transférés dans l'ordinateur du responsable du traitement informatique.

Transformation de fichiers CSPRO en fichiers SPSS

Le logiciel CSPRO utilisé pour la saisie des données permet d'obtenir un fichier sous format ASCII. Les fichiers de ce format ne sont pas directement utilisables par un programme SPSS, logiciel retenu pour l'apurement des données. Même si on voulait lire directement le fichier produit par CSPRO, Il y aurait d'ailleurs une difficulté supplémentaire. SPSS ne peut lire efficacement que les fichiers plats (un enregistrement correspond à un individu statistique) ou les fichiers rectangles (un nombre fixe d'enregistrements correspondent à un individu). Or, dans le cas d'espèce, le fichier produit comporte trois types d'enregistrements : un enregistrement ménages (variables H, E et P), autant d'enregistrements socio-démographiques (variables M) qu'il y a de personnes dans le ménage et autant d'enregistrements emploi (variables EA, AP, AS, R, C, TP et RHA) qu'il y a de personnes de 10 ans et plus dans le ménage. Ce fichier n'est pas plat (car il y a plus d'un enregistrement par ménage) et il n'est pas rectangle (car le nombre de personnes varie d'un ménage à l'autre). Il faut donc utiliser des procédures spécifiques pour transformer le fichier CSPRO en un fichier qui peut être traité par SPSS.

Dans un premier temps, on transforme les fichiers CSPRO (LOT01A à L0T01J pour la première série de correction, LOT02A à L0T02J pour la deuxième, etc.) en trois fichiers rectangles. Cette opération se fait à l'aide de la procédure « EXPORT ». Pour la mettre en î uvre, après avoir lancé le programme de saisie (icône ***SAISIE Phase1***), on va au menu « ***TOOLS*** » et on choisit le sous-menu « ***EXPORT DATA*** ». La procédure est automatique.

i) Le logiciel vous demande simplement d'ouvrir un dictionnaire ; vous choisissez le fichier « parsta~1.dcf » dans le répertoire ENQ123\Phase1\PRGM.
ii) Vous devez ensuite choisir les variables à exporter ; en choisissant le questionnaire entier, vous choisissez d'exporter toutes les variables.
iii) Vous choisissez ensuite le menu « ***FILE*** » et vous faites « ***RUN*** » ; le logiciel vous demande ensuite le nom du fichier à exporter.
iv) Vous allez au répertoire approprié (normalement C:\ENQ123\ PHASE1\DATA) et vous choisissez les dix fichiers à exporter (on le fait en maintenant la touche « ***Ctrl*** » appuyée ï opération classique dans Windows) : LOT01A à L0T01J ; ces dix fichiers sont implicitement consolidés lors de cette opération.

Le résultat est la création de trois fichiers plats : un fichier ménage contenant les variables H, E et P et deux fichiers individus dont l'un composé des variables socio-démographiques des membres du

ménage et le second des variables sur l'activité des membres du ménage. ***Ces fichiers portent implicitement les noms HABITAT.txt, DEMO.txt et EMPLOI.txt***. Ils ont deux caractéristiques : premièrement, ils disposent de séparateurs et la version de SPSS utilisée permet de les lire automatiquement sans avoir à écrire un programme compliqué ; deuxièmement, les noms des variables sont consignés à la première ligne et peuvent être conservés au moment de la lecture. Pour lire chacun de ces fichiers, dès lors que vous êtes dans le logiciel SPSS, il suffit de choisir le menu « ***FILE*** » et le sous-menu « ***READ TEXT DATA*** ». ***Vous pouvez sauvegarder chacun d'eux avec le même nom que précédemment (on propose quelques variantes par la suite)***, mais avec l'extension « SAV » ; cette extension est d'ailleurs attribué automatiquement aux fichiers systèmes SPSS.

Sauvegarde des fichiers sous format SPSS

Pour faciliter l'apurement des données, il est préférable de disposer de deux fichiers seulement pour la phase 1[1]. Le premier fichier est le fichier ménage, il est obtenu à l'aide de l'exécution de la procédure « ***READ TEXT DATA*** » précédente, aucune opération supplémentaire n'est donc nécessaire. Le second est le fichier individu. Pour l'obtenir, il convient de fusionner les fichiers DEMO.SAV et EMPLOI.SAV ; le programme ESI01a.SPS vous permet d'obtenir un fichier total résultat de la consolidation des deux fichiers précédents.

Dénomination des fichiers

Pour suivre de manière rigoureuse le processus d'apurement des données, il convient d'adopter des règles de dénomination des fichiers. Il est proposé de traiter l'enquête en 5 vagues différentes, numérotées de 1 à 5. Pour chacune des vagues, on peut procéder à plusieurs phases de correction. Il est important de conserver les fichiers de chacune des phases.

Les fichiers SPSS bruts du premier batch de la phase 1 seront appelés respectivement HABITA1a.SAV et INDIVI1a.SAV. D'après les instructions précédentes, le fichier HABITA1a.SAV est obtenu directement à la suite de l'exécution de la procédure « ***READ TEXT DATA*** » de SPSS à partir du fichier de départ HABITAT.txt. Quant au fichier INDIVI1a.SAV, il est obtenu à la suite de l'exécution du programme intitulé ESI01a.SPS, programme qui permet le fusion des fichiers DEMO.SAV et EMPLOI.SAV ; cette fusion peut également se faire automatiquement.

Des programmes d'apurement sont exécutés sur ces premiers fichiers et une liste des erreurs est produite. Si le taux d'erreur est trop élevé (à l'appréciation du Directeur technique de l'enquête ou de son adjoint, il faudrait vérifier s'il s'agit des erreurs de saisie, auquel cas il faut reprendre la saisie ; ou des erreurs de terrain, auquel cas il faut un retour de terrain. Si le taux d'erreur est acceptable, on procède aux corrections sur CSPRO dans les fichiers LOT01x (x = A, é , J). Après les premières corrections, on reprend la procédure d'exportation du fichier pour aboutir à des fichiers SPSS dénommés HABITA1b.SAV et INDIVI1b.SAV, etc. Il faut noter que les questionnaires non rejetés lors du traitement des fichiers HABITA1a et INDIVI1a ne devraient plus l'être dans la suite.

Procédure d'apurement

Deux approches sont possibles en matière d'apurement des données. La correction automatique suppose de prévoir des imputations à toute erreur décelée sans avoir à consulter les données de base (questionnaires). Les imputations se font à l'aide des procédures telles que le « Cold-deck », le « Hot-deck », des procédures économétriques, etc. Cette approche est souvent utilisée dans les recensements de population. L'autre approche, celle en général adoptée dans les enquêtes est la correction semi-automatique. Il s'agit de déceler les erreurs, de les lister et de faire un retour dans les questionnaires pour procéder aux corrections ; c'est la procédure adoptée dans le cadre de cette opération.

Les programmes d'apurement qui sont rédigés ne prévoient pas les tests d'amplitude sur les données. Pour ce faire et avant toute chose, ***il convient de faire des tris à plat systématique sur toutes les variables du fichier***. Ces tris à plat permettent d'épingler les variables où il y aurait des modalités « out of range », on écrit alors une procédure simple pour lister les individus qui auraient ces

[1] Afin de faciliter le traitement des données au niveau régional, plusieurs fichiers de taille plus modeste (en fait un fichier par module du questionnaire) seront créés après la phase d'apurement.

modalités. En outre, le fait de réaliser ces tris à plat permet d'avoir une première idée sur la qualité des données.

Les programmes d'apurement se nomment ESI02.SPS, ESI03.SPS, é , ESI14.SPS. Chacun de ces programmes permet d'obtenir la liste des erreurs d'un module du questionnaire. La logique de ces programmes est la même et elle est relativement simple. Après la lecture des fichiers, une première partie du programme permet de détecter les erreurs et la seconde partie de les lister.

Quand on a obtenu la liste des erreurs, les corrections ne se font pas dans le fichier SPSS, mais plutôt dans le fichier consolidé CSPRO. Pour prendre un exemple concret, on obtient 3 fichiers SPSS à l'aide du fichier LOT01 ; après exécution du programme ESI01a, on obtient les deux fichiers HABITA1a.SAV et INDIVI1a.SAV. On exécute les programmes ESI02.SPS à ESI14.SPS sur ces deux fichiers et on obtient la liste des erreurs. A l'aide de cette liste, on rentre dans CSPRO pour corriger les fichiers LOT01x. Après l'exécution des programmes, on conserve les fichiers HABITA1a.SAV et INDIVI1a.SAV (fichiers bruts ; surtout ne pas les détruire). Pour la deuxième série de correction (on en est toujours à la première vague), on crée de nouveaux fichiers HABITA1b.SAV et INDIVI1b.SAV et on reprend la procédure précédente. Cette procédure est exécutée autant de fois que nécessaire jusqu'à l'obtention de fichiers propres. Les fichiers intermédiaires (entre les fichiers bruts et les fichiers apurés) sous format SPSS sont également archivés quelque part ; en tout cas il serait imprudent de les détruire avant la fin des travaux d'analyse. Pour la deuxième vague de corrections (500 prochains ménages), les fichiers CSPRO s'appellent LOT02x, les fichiers SPSS, HABITA2a.SAV et INDIVI2a.SAV, etc.

Pour ce qui est des corrections proprement dit, la décision n'appartient surtout pas aux informaticiens et encore moins aux agents de saisie ; le responsable technique doit superviser l'apurement des données. En fait quand il s'agit d'une erreur de saisie, le retour au questionnaire permet de corriger l'erreur. Par contre, pour les erreurs de terrain, la réponse n'est pas toujours aisée. Mais il faut utiliser l'information disponible pour redresser les incohérences. A titre d'exemple, un individu qui déclare ne pas avoir de revenu de transfert (RHA5a = 2) mais qui donne un montant correspondant (RHA5b # Blanc), tout laisse à penser qu'il faut changer RHA5 en 1. Evidemment toutes les erreurs ne peuvent être corrigées.

Une procédure spéciale de correction automatique des revenus est prévue, elle ne sera exécutée qu'à la fin des autres corrections comme préalable aux travaux d'analyse.

Sauvegarde des fichiers définitifs

A la fin des opérations d'apurement, on dispose de deux fichiers pour chaque batch : un fichier contenant les informations des modules H, E et P et un fichier sur les caractéristiques des personnes (caractéristiques socio-démographiques et sur l'activité des membres des ménages). Evidemment, le fichier le plus intéressant est le fichier entièrement corrigé (le dernier fichier du lot). Il s'agit donc d'additionner ces différents fichiers pour obtenir les 2 fichiers définitifs de l'enquête, fichiers utilisés par l'analyse***. Le programme ES101b.SPS permet d'obtenir ce fichier définitif. Ce programme est donc le dernier à être exécutée puisqu'il ne l'est qu'après toutes les corrections***. L'hypothèse faite lors de l'écriture de ce programme est que chaque LOT est corrigé trois fois, la version définitive des fichiers d'un lot est donc la version « d » ; chaque pays va procéder aux adaptations appropriées.

Un aspect important pour le stockage des fichiers définitifs est leur documentation ; les variables et les modalités des variables doivent avoir des labels, ce qui n'est pas le cas pour les fichiers disponibles jusqu'alors, le programme ESI01b.SPS corrige cet état de fait.

Les fichiers définitifs auront les intitulés devant permettre de s'y retrouver facilement. Par exemple, le fichier « Habitat » de la phase 1 du Bénin : ESI1BEH1 et le fichier « Individu » de la phase 1 du Bénin : ESI1BEI1. ESI est mis pour enquête secteur informel (plus court que E123 qui était mieux), 1 qui suit pour la phase 1, BE pour le Bénin, H ou I respectivement pour habitat et individu et le dernier 1 pour l'année 2001.

Remarques finales

Les programmes sont rédigés sur la base des questionnaires du Mali, pour chaque pays ils demandent donc à être adaptés. Les adaptations seront surtout relatives au programme ESI01b (modalités des variables en particulier) et aux programmes d'apurement concernant la fiche ménage (ESI02.SPS à ESI06.SPS), c'est à ce niveau que les différences entre les pays sont les plus nombreuses. Ces différences concernent en particulier les variables relatives à l'éducation. Au mali, le premier cycle du secondaire se fait en 3 ans alors qu'il se fait en 4 ans dans les autres pays. Les tests de cohérence où intervient la variable M15 doivent donc être revus.

D'autres adaptations (relativement simples) concernent les noms des fichiers en lecture. En outre, les erreurs de syntaxe ne sont pas à exclure. En définitive, les informaticiens doivent s'imprégner de ces programmes avant de les exécuter.