

Les différents types de fichiers nécessaires à l'analyse des données de CP en Côte d'Ivoire pour l'année scolaire 1995-1996 sont rangés dans des répertoires spécifiques :

- C:\Cd_pasec\CI\1995-96\niveau2\data\mdb : Ce répertoire contient la base de données ACCESS CI_9596.mdb constituée dans la première partie de ce chapitre
- C:\Cd_pasec\CI\1995-96\niveau2\data\txt : Ce répertoire contient les fichiers de données au format texte (extension .txt) issus de l'exportation des différentes tables de la base ACCESS
- C:\Cd_pasec\CI\stata\do\cohorte : Ce répertoire contient les fichiers programme STATA (extension .do)
- C:\Cd_pasec\CI\stata\dta : Ce répertoire contient les fichiers de données transformées par les programmes STATA (extension .dta)
- C:\Cd_pasec\CI\stata\log : Ce répertoire contient les fichiers de résultat (extension .log)

Examinons maintenant le schéma qui permet de passer des données brutes issues de la base ACCESS «CI_9596.mdb» à des fichiers de données transformées prêtes pour l'analyse (voire graphique «schéma d'analyse principal», page suivante).

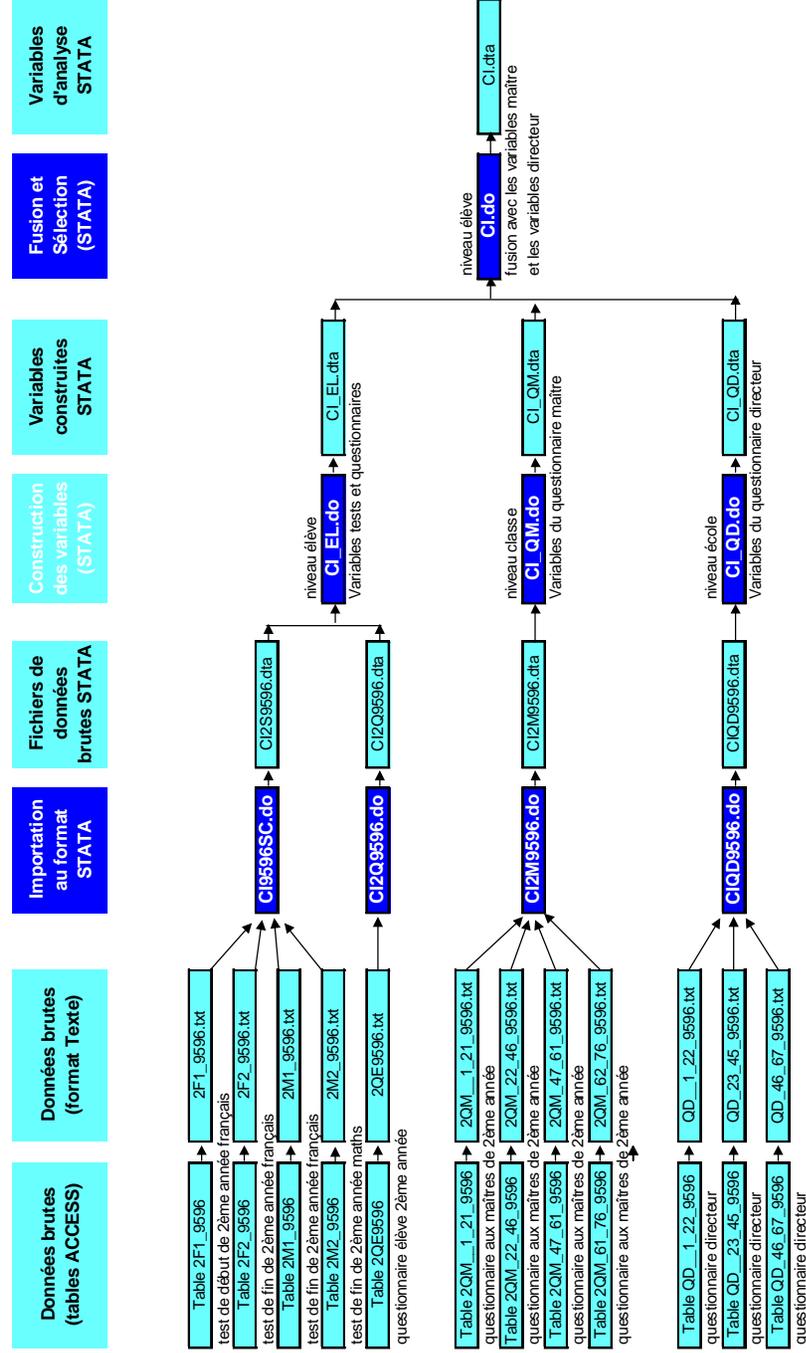
Dans le graphique, les programmes sont foncés, et les données sont en clair.

Suivons le processus de création qui nous mène de la base ACCESS au fichier qui nous servira pour mener les analyses (fichier CI.dta sur la droite du graphique).¹

La première colonne représente les tables de données à l'intérieure de la base ACCESS «CI_9596.mdb».

¹ Ce schéma est une simplification du schéma présent sur le CD-PASEC, puisque ce dernier inclut des données pour plusieurs années (d'où le suffixe 9596 présent dans les noms de fichiers, et que nous n'avons pas introduit jusqu'alors).

PASEC Côte d'Ivoire, Données 1995-1996, Schéma d'analyse principal



Note: L'exécution en série de ces programmes s'obtient en lançant \C:\stata\d\c\chorel\CI_batch.do

La première opération consiste à exporter ces tables ACCESS dans le répertoire C:\Cd_pasec\CI\1995-96\niveau2\data\txt . Ceci se fait depuis la base ACCESS, table par table, avec la fonction d'exportation incluse dans le logiciel. Parmi les différents formats proposés, nous choisirons le plus universel, le format texte. Certaines options devront être précisées, comme celle de séparation «déliimitée» entre les champs (par opposition à «longueur fixe»), ainsi que le choix du caractère «virgule» comme séparateur, et du caractère «point» comme symbole décimal, pour se conformer aux habitudes «anglo-saxonnes» du logiciel STATA. Les 12 tables de données doivent ainsi être exportées (nous laissons de côté les tables ECOLES, CLASSES, et ELEVES, qui ont déjà joué leur rôle d'intégration des différents niveaux, et dont les renseignements nominatifs NOMEcole, NOMMAITRE et NOMELEVE ne nous intéressent pas dans l'analyse).

Une fois les douze tables de données d'ACCESS transformées en autant de fichiers de données au format .txt (deuxième colonne du schéma) , nous pouvons passer à l'élaboration des programmes STATA chargés de traduire toutes ces données au format STATA. Nous créerons quatre programmes à cette fin : un chargé des données de test aux élèves, un chargé des données des questionnaires élèves, un chargé des données du questionnaire aux maîtres, et un chargé du questionnaire aux directeurs.

Voici, à titre d'illustration, certains moments «forts» de ces programmes¹, à partir de l'exemple de «CI2Q9596.do», chargé, comme son nom l'indique avec une concision digne d'éloges, de l'importation au format stata des données de questionnaire élève en deuxième année (CP), en Côte d'Ivoire en 1995-1996.

Voici donc quelques extraits choisis de CI2Q9596 :

Ouverture d'un fichier LOG chargé de rendre compte de l'exécution du programme :

```
log using c:\CD_PASEC\CI\stata\log\CI2Q9596.log, replace
```

¹Tous ces fichiers peuvent être lus dans un éditeur de texte (type Bloc-note) ou un logiciel de traitement de texte (type Word), à partir du CD-ROM PASEC.

Importation des données au format texte, en même temps qu'attribution d'un nom STATA pour les différentes variables :

```
infile NUMECOLE NUMCLASS NUMELEVE /*
*/ QE9596_A QE9596_B QE9596_C QE9596_D QE9596_E QE9596_F QE9596_G /*
*/ QE9596_H QE9596_I QE9596_J QE9596_K QE9596_L QE9596_M QE9596_N /*
*/ QE9596_O QE9596_P QE9596_Q QE9596_R QE9596_S QE9596_T QE9596_U /*
*/ QE9596_V QE9596_W QE9596_X QE9596_Y QE9596_Z QE9596AA QE9596AB /*
*/ QE9596AC QE9596AD QE9596AE QE9596AF QE9596AG QE9596AH QE9596AI /*
*/ QE9596AJ QE9596AK QE9596AL QE9596AM QE9596AN QE9596AO QE9596AP /*
*/ QE9596AQ QE9596AR QE9596AS QE9596AT QE9596AU QE9596AV QE9596AW /*
*/ QE9596AX /*
*/ using c:\CD_PASEC\CI\1995-96\niveau2\data\txt\2QE9596.txt
```

Sauvegarde des données du questionnaire élève au format STATA :

```
save c:\CD_PASEC\CI\stata\dta\CI2Q9596.dta, replace
```

Nos données sont donc maintenant disponibles, dans leur intégralité, au format STATA, sous la forme de quatre fichiers (CI9596SC.dta pour les scores des élèves, CI2Q9596.dta pour les questionnaires élèves, CI2M9596.dta pour les questionnaires maîtres, et CIQD9596.dta pour les questionnaires directeurs). A noter que contrairement à ACCESS, STATA accepte un grand nombre de variables à l'intérieur d'une même «table» (fichier .dta), ce qui permet, par exemple, de rassembler les quatre tables ACCESS consacrées au questionnaire maître au sein d'un seul fichier de données STATA.

Cette fusion de plusieurs fichiers de même niveau en un seul s'effectue de manière simple. Voici les instructions pour la fusion des quatre fichiers STATA provisoires pour le questionnaire maître en un seul fichier STATA dans le programme CI2M9596.do :

```
clear

use C:\CD_PASEC\CI\stata\dta\CIQM_121.DTA

sort NUMECOLE NUMCLASS
merge NUMECOLE NUMCLASS using C:\CD_PASEC\CI\stata\dta\CIQM2246.dta
drop _merge
sort NUMECOLE NUMCLASS
merge NUMECOLE NUMCLASS using C:\CD_PASEC\CI\stata\dta\CIQM4761.dta
drop _merge
sort NUMECOLE NUMCLASS
merge NUMECOLE NUMCLASS using C:\CD_PASEC\CI\stata\dta\CIQM6276.dta
drop _merge
```

```

sort NUMECOLE NUMCLASS

! del C:\CD_PASEC\CI\stata\dta\CIQM_121.dta
! del C:\CD_PASEC\CI\stata\dta\CIQM2246.dta
! del C:\CD_PASEC\CI\stata\dta\CIQM4761.dta
! del C:\CD_PASEC\CI\stata\dta\CIQM6276.dta

save C:\CD_PASEC\CI\stata\dta\CI2M9596.dta, replace

```

Nous voyons, que comme dans ACCESS, les deux champs clés d'identification au niveau classe (NUMECOLE et NUMCLASSE) sont mis à contribution pour mettre bout à bout les lignes correspondant à un même maître dans les différents questionnaires.

Ces données sont encore brutes : il reste à y construire des variables pour l'analyse (scores, variables politiques, variables contextuelles, etc). Pour cela, nous créons trois programmes (un pour les données niveau élève, un pour les données niveau classe, et un pour les données niveau maître).

Ces programmes effectuent plusieurs tâches préliminaires avant de construire chaque variable :

- recherche et élimination (ou correction) des valeurs absurdes
- imputation (éventuellement) des valeurs manquantes
- construction de la variable

Les principes à observer dans cette phase sont détaillés au chapitre suivant. Indiquons juste un exemple concernant le calcul du score pour le pré-test de français, à partir du résultat aux différents items (codés 0 pour faux et 1 pour juste) :

```

/* *****
                                     CALCUL DU SCORE PRE-TEST FRANCAIS
*****
*/

gen SINI2F=INI2F__A+INI2F__B+INI2F__C+INI2F__D+INI2F__E+INI2F__F+ /*
*/   INI2F__G+INI2F__H+INI2F__I+INI2F__J+INI2F__K+INI2F__L+ /*
*/   INI2F__M+INI2F__N+INI2F__O+INI2F__P+INI2F__Q+INI2F__R+ /*
*/   INI2F__S+INI2F__T+INI2F__U+INI2F__V+INI2F__W+INI2F__X+ /*
*/   INI2F__Y

```

ou bien encore comment créer la variable NIVEAUVI à partir de la possession de un ou de plusieurs objets :

```

/* *****
INDICATEUR DE NIVEAU DE VIE
*****
*/

gen FAUTEUIL=QE9596_G==1 if QE9596_A~= .
gen FRIGO=QE9596_I==1 if QE9596_A~= .
gen ROBINET=QE9596_J==1 if QE9596_A~= .
gen LIT=QE9596_K==1 if QE9596_A~= .
gen ELECTRIC=QE9596_L==1 if QE9596_A~= .
gen LAMPETR=QE9596_M==1 if QE9596_A~= .
gen VOITURE=QE9596_N==1 if QE9596_A~= .
gen VELO=QE9596_O==1 if QE9596_A~= .
gen MOBYLETT=QE9596_P==1 if QE9596_A~= .
gen CHARENTE=QE9596_Q==1 if QE9596_A~= .
gen VIDEO=QE9596_R==1 if QE9596_A~= .
gen TV=QE9596_S==1 if QE9596_A~= .
gen RADIO=QE9596_T==1 if QE9596_A~= .
gen CUISIGAZ=QE9596_U==1 if QE9596_A~= .
gen EAUTOIL=QE9596_W==1 if QE9596_A~= .
gen CHARRUE=QE9596_X==1 if QE9596_A~= .

gen NIVEAUVI=VIDEO+VOITURE+FRIGO if QE9596_A~= .

```

En fin de programme, on se débarrasse des variables brutes, pour ne pas surcharger les fichiers de données de variables brutes inutiles. Voici comment faire pour effacer les variables brutes du questionnaire élève

```
drop QE*
```

Toutes les variables brutes commençant par QE (par exemple, QE_1A) sont aussitôt effacées.

Les trois fichiers de données issus de ce traitement sont représentés dans la sixième colonne de notre schéma (CI_EL.dta pour les variables construites niveau élève, CI_QM pour les données de niveau classe, et CI_QD pour les variables de niveau école).

Il reste alors à fusionner ces trois fichiers au sein d'un seul, de niveau élève. C'est le rôle du fichier programme CI.do. Cette fusion entre fichiers de différents niveau s'effectue en jouant avec les champs clés NUMECOLE, NUMCLASSE et NUMELEVE :

```
use C:\CD_PASEC\CI\stata\dta\CI_QM.dta

sort NUMECOLE NUMCLASS

/* APPORT DES QUESTIONNAIRES ELEVES */

merge NUMECOLE NUMCLASS using C:\CD_PASEC\CI\stata\dta\CI_EL.dta
tab _merge

/* suppression des questionnaires maitres sans eleves correspondants */
drop if NUMELEVE==.

drop _merge

/* APPORT DES QUESTIONNAIRES DIRECTEURS */

sort NUMECOLE
merge NUMECOLE using C:\CD_PASEC\CI\stata\dta\CI_QD.dta
```

Une fois sauvegardées sous la forme d'un fichier de données STATA (CI.dta), ces données sont désormais disponibles pour l'analyse :

Ainsi, en mode interactif, l'exécution des deux commandes suivantes :

```
use C : \CD_PASEC\CI\stata\dta\CI.dta
reg STFIN2FM STINI2FM
```

nous donnera la part de variance du score de fin d'année expliquée par le score de début d'année. Mais ce serait déjà anticiper sur l'analyse...