Analyse technique des données en démarche préalable.

La collecte des informations variant nécessairement au fil des années, des éléments ont été ajoutés ou supprimés. Il n'est donc pas possible d'utiliser des éléments dont la présence n'est pas permanente sur toute la période de collecte.

Afin de déterminer quels éléments sont communs à chaque fichier, quelques manipulations techniques ont été nécessaires.

Sortir d'un format spécifique.

Les fichiers fournis sont au format Stata.

La première opération réalisée consiste à les exporter de Stata vers un format plus aisément manipulable, le format CSV. Stata permet un fonctionnement en mode « batch », cela a donc été utilisé. Chaque fichier Stata s'est vu exporté vers un fichier CSV.

L'opération a été effectuée deux fois ; une fois pour connaitre les variables communes, une seconde fois pour n'exporter que ces variables communes.

Quelles sont les variables communes ?

Au premier export, chaque fichier CSV a été construit avec les noms des variables sur la première ligne, et doté d'un nombre minimal de ligne, soit une ligne.

Un logiciel libre, en Python, disponible sous licence GPL, a été adapté pour déterminer, à partir de cet ensemble de fichier, quelles sont les colonnes / variables présentes dans chacun des fichiers.

Cette liste de variables est utilisé plus loin.

Maintenant que nous connaissons quelles variables exporter, nous recommençons l'export, mais en précisant quelles variables nous voulons.

Une simple concaténation assemble l'ensemble des fichiers produits en un seul, que nous importons dans Stata.

Peine perdue, Stata 32 bits ne peut gérer un tel volume de données.

Il va falloir être plus sélectif encore. Un examen superficiel des informations et, surtout, la lecture du dictionnaire de données, fait apparaitre que certaines lignes ne sont pas nécessaires.

Sélectionner par conditions

La solution pour faire une sélection conditionnelle, et pouvoir aisément en faire une autre si je décidais de changer de critères ; la solution passe par l'utilisation d'une base de données.

Le fichier CSV a donc été importé dans une base de données (SQLite), puis le résultat d'une requête a été exporté dans un nouveau fichier.

Enfin, ce fichier a été importé avec succès dans Stata.