LE LOGIEL POULPE : MODÉLISATION INFORMATIQUE

Jean-Noël Petit

Ce logiciel s'adresse aux responsables d'enquête soucieux d'évaluer la qualité des données recueillies sur des échantillons tirés de sondages complexes, à plusieurs degrés et plusieurs phases. Il apporte une estimation de l'erreur due à l'échantillonnage, en estimant la variance de l'estimateur d'Horvitz-Thompson sur les variables d'intérêt, et fournit l'intervalle de confiance à 95% centré sur le total pondéré.

Développé à partir du langage Sas Macro, il devrait être disponible prochainement sur les sites Mvs de l'Insee, et dans l'environnement Sas Windows, dans une version de pré-production.

1. Sur quelles bases reposent les calculs ?

Le logiciel s'appuie sur les informations issues de 3 sources différentes :

- le fichier des données résultant de l'enquête (appelé DATA), contrôlées et corrigées, éventuellement redressées à partir de données exogènes, par un logiciel approprié (par exemple Calmar),
- le fichier décrivant le plan de sondage (appelé MODELE),
- un fichier auxiliaire, ou fichier géographique (appelé GEO), avec les effectifs des unités administratives ou statistiques (appelées entités géographiques), destinés au calcul des probabilités d'inclusion.

Ces probabilités d'inclusion, indispensables au calcul des variances des estimateurs, figurent rarement dans le fichier de l'utilisateur, et doivent donc être évaluées par le logiciel, pour les différents degrés du sondage, à partir de :

- n : la taille de l'échantillon, présent implicitement dans le fichier de données,
- N : la taille de la population dans laquelle on a effectué le tirage : elle ne figure généralement pas dans le fichier de données, mais dans une source auxiliaire (GEO) ; cette source donne les effectifs de toutes les entités géographiques intervenant dans le tirage,

- le type de tirage : aléatoire simple, systématique, proportionnel à la taille, exhaustif,...

- des données auxiliaires, par exemple la taille pour les sondages proportionnels à la taille, ou les variables de tri pour les sondages systématiques...

Les formules utilisées permettent de s'affranchir des probabilités d'inclusion d'ordre 2, du type π_{ij}.

Pour ce calcul, le logiciel rapproche les sources d'information DATA et GEO à partir des identifiants des unités tirées, qui doivent donc être présents dans ces 2 sources. Il peut y avoir là un travail préparatoire à effectuer pour harmoniser les identifiants de ces sources, dans leurs dénominations et leurs types.

Pour les enquêtes ménages tirées de l'échantillon maître du RP 90, l'unité Méthodes Statistiques a constitué le fichier géographique de toutes les unités présentes dans ce sondage complexe : départements, communes, régions, strates unités primaires... avec leurs effectifs.

Une fois déterminées les probabilités d'inclusion, on effectue le calcul des variances de l'estimateur de Horvitz-Thompson, en s'appuyant sur un modèle de représentation (MODELE) apte à étendre à plusieurs degrés les formules connues pour les sondages à 1 degré.

2. Le modele sous-jacent au logiciel : comment represente-t-on un sondage a plusieurs degres ?

Lorsque l'on s'intéresse au calcul de la variance de l'estimateur de Horvitz-Thompson, on dispose d'un arsenal de formules pour un tirage à 1 degré, en fonction du type de tirage.

Exemple 1 : sondage à 1 degré proportionnel à la taille

Dans l'exemple ci-dessous, on tire des communes dans un canton proportionnellement à la taille de la commune.

On estime la somme d'une variable 'au niveau commune' y au niveau du canton par l'estimateur d'Horvitz-Thompson :

$$\hat{t} = \sum_s \frac{y_k}{\pi_k}$$

On connaît l'estimateur de la variance de cette somme, qui vaut :

$$\hat{V} = \frac{n}{n-1} \sum_k (1-\pi_k)\left(\frac{y_k}{\pi_k} - \sum_s a_k \frac{y_k}{\pi_k}\right)^2$$

avec : $a_k = \frac{(1-\pi_k)}{\sum_s (1-\pi_k)}$ (π_i : probabilité d'inclusion)

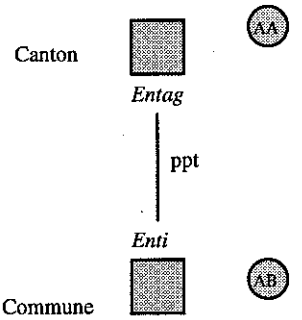
π_i = n* (Taille *de l'entité tirée) / Σ(Taille des entités tirables)*

soit : π_i = n (Effectif de la commune) / (Effectif du canton)*

On représente dans le logiciel un tel sondage par un arc sur lequel figurent ces informations (type de sondage, noms des entités), et on identifie les extrémités de l'arc par un code qui permettra de faire le lien entre les sondages élémentaires successifs.

On tire des communes dans des cantons proportionnellement à leur taille. On désigne par les termes :

- *entité tirée* : la commune (Enti),
- *entité d'agrégation* : le canton (Entag),
- *type de tirage* : tirage à probabilités proportionnelles à la taille (ppt).



On associe à chaque sondage un ensemble de données nécessaires au calcul des estimateurs :

- le type de tirage élémentaire : aléatoire simple, systématique,...
- les variables utilisées pour le tirage (par exemple la taille pour les tirages à probabilités proportionnelles à la taille),

- des données annexes relatives au sondage, et intervenant dans les formules : taux de sondage, population totale,... ; certaines de ces données sont fournies à part, d'autres (par exemple la population de l'échantillon) sont évaluées à partir de la base des données (DATA),
- entité tirée (ENTI) (canton, commune, groupe de communes, logement,...) ; c'est au niveau de cette entité que l'on applique les formules d'estimation liées aux éventuels sondages ultérieurs,
- entité d'agrégation (ENTAG) dans laquelle on tire les entités ENTI ; c'est à cette entité (ENTAG) que l'on applique les fonctions d'estimation,
- les variables d'intérêt sur lesquelles on applique les formules.

Exemple 2 : sondage à 2 degrés

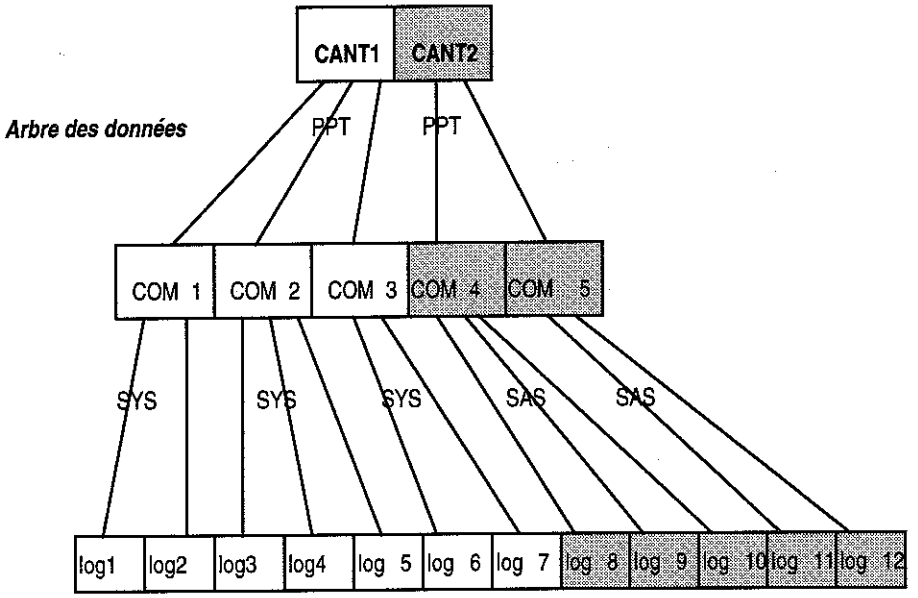
Dans ce plan de sondage, on considère un arrondissement composé de 2 cantons :
- le canton 1, dans lequel on tire 3 communes : COM1, COM2, COM3
- le canton 2, dans lequel on tire 2 communes : COM4, COM5.

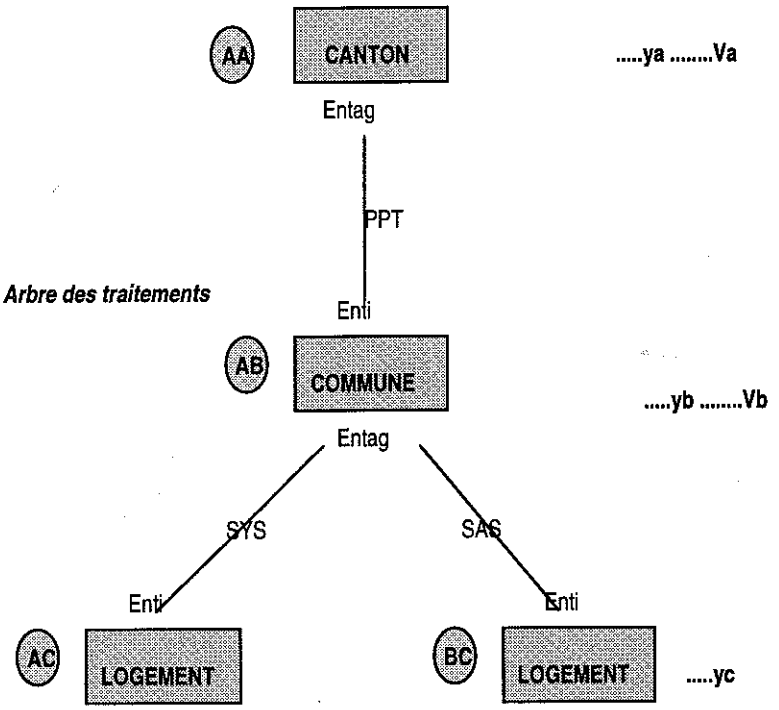
Puis, dans les communes du canton 1, on tire des logements par un tirage systématique (logements LOG1 à LOG7), dans celles du canton 2, des logements par un tirage aléatoire simple (logements LOG8 à LOG12).

Pour modéliser ce nouveau sondage, à 2 degrés, on étend le modèle précédent en y adjoignant 2 arcs, nommés AB-AC et AB-BC, pour obtenir finalement un «arbre» composé de :

- 3 arcs (AA-AB, AB-AC, AB-BC),
- 2 noeuds (AA et AB) et 2 feuilles (AC et BC).

Pour l'arc AA-AB, l'entité tirée est la commune, l'entité d'agrégation est le canton ; pour les arcs AB-AC et AB-BC, ce sont le logement et la commune.


Arbre des données
CANT1
CANT2
PPT
PPT
COM 1
COM 2
COM 3
COM 4
COM 5
SYS
SYS
SYS
SAS
SAS
log1
log2
log3
log4
log 5
log 6
log 7
log 8
log 9
log 10
log 11
log 12


AA
CANTON
.....ya .....Va
Entag
PPT
Arbre des traitements
Enti
AB
COMMUNE
.....yb .....Vb
Entag
SYS
SAS
Enti
Enti
AC
LOGEMENT
BC
LOGEMENT
.....yc

On désigne par :

* y_c les valeurs de la variable d'intérêt y au niveau des feuilles AC et BC,

* y_b et V_b les variables " estimateur de total " et "estimateur de variance " au niveau du noeud AB,

* y_a et V_a les variables " estimateur de total " et "estimateur de variance " au niveau du nœud AA.

Lors du traitement des arcs AB-AC et AB-BC, on évalue les estimateurs yb et Vb en fonction des yc. On traite ensuite l'arc AA-AB en calculant les estimateurs y_a et V_a en fonction des y_b et des V_b.

Cette évaluation met en jeu des formules qui dépendent du type de tirage. Les y_b et V_b une fois estimés, deviennent les variables d'intérêt pour le sondage du niveau supérieur, et on renouvelle le processus jusqu'à aboutir à la racine de l'arbre (inversé).

Sur cet exemple, on obtient ainsi :

$$y_{com1} = f_{sys}(y_{lg1}, y_{lg2})$$

$$V_{com1} = g_{sys}(y_{lg1}, y_{lg2})$$

$$y_{com4} = f_{sas}(y_{lg8}, y_{lg9}, y_{lg10})$$

$$V_{com4} = g_{sas}(y_{lg8}, y_{lg9}, y_{lg10})$$

$$y_{cant1} = f_{ppt}(y_{com1}, y_{com2}, y_{com3})$$

$$V_{cant1} = g_{ppt}(y_{com1}, y_{com2}, y_{com3}) + g_{ppt} * (V_{com1}, V_{com2}, V_{com3})$$

Les fonctions f_{sys} , g_{sys} ..., f_{ppt} , g_{ppt} et g^*_{ppt} qui représentent des formules de calcul appropriées à chaque type de tirage, sont la traduction de la formule de Raj :

$$\hat{V}(\hat{Y}) = f(\hat{t}) + \sum_s w_{is} \hat{V}_i$$ (se reporter à l'article de N. Caron)

De par les propriétés de l'estimateur de Horvitz-Thompson, elles s'expriment dans le logiciel par la formulation suivante :

la variance au niveau p est la somme de 2 termes :

- *la variance des sommes obtenues au niveau p-1,*
- *la somme pondérée des variances obtenues au niveau p-1, en utilisant les formules du sondage permettant de passer du niveau p au niveau p-1.* (p=1 pour les feuilles et p=n pour la racine de l'arbre).

Généralisation à n degrés

On peut étendre à n degrés le modèle retenu pour les sondages à 2 degrés, en reliant les divers arcs des sondages élémentaires : on aboutit ainsi à l'arbre modélisant le sondage à plusieurs degrés (voir en annexe 1 à titre d'exemple, une branche de l'arbre modélisant l'échantillon maître 1990 pour les communes de moins de 20 000 hab.)..

Le calcul s'exerce sur les données du fichier pour produire les premiers estimateurs (en général estimateur du total et de la variance pour n variables d'intérêt) relatifs aux derniers sondages élémentaires. Ces estimateurs deviennent les variables d'intérêt du sondage supérieur représenté aussi par un arc, avec des entités tirées qui sont les entités d'agrégation du sondage inférieur.

Le calcul récursif est conduit dans l'ordre chronologique inverse des opérations de tirage des entités : on commence par traiter les niveaux inférieurs (c'est-à-dire les dernières entités tirées) pour remonter ensuite aux niveaux supérieurs.

Une stratification est assimilée à un tirage avec un taux de sondage égal à 1.

3. Les différentes étapes

Après une phase préparatoire de mise en cohérence des sources, les traitements sont groupés en deux étapes.

Un premier ensemble d'étapes au cours desquelles :

- l'utilisateur décrit le plan de sondage au niveau de chaque arc, sous la forme d'une table SAS (appelée MODELE) ; cette modélisation du sondage demande une parfaite connaissance du plan de sondage ;

- le logiciel calcule les probabilités d'inclusion élémentaires (c'est-à-dire relatives à un sondage élémentaire), lorsqu'elles sont absentes du fichier de données, et globales (résultant des différents tirages successifs), à partir d'une table SAS

(appelée GEO) dans laquelle on aura inscrit au préalable les populations de toutes les unités tirées (ou leur taille pour les sondages à probabilités proportionnelles à la taille).

Cette étape est réalisée une seule fois.

Un deuxième ensemble d'étapes pour l'application des formules :

- on précise les statistiques d'intérêt : totaux de variables ou statistiques complexes,
- le logiciel calcule les variances estimées de ces statistiques.

Cette étape est relancée à chaque fois que l'on étudie de nouvelles statistiques. Elle fait appel à un ensemble de modules lancés individuellement ou générés à partir de noms d'étape globale : ESTIVAR (pour l'estimation des variables) ou ESTIFON (pour l'estimation de statistiques complexes).

On lance toutes les étapes à partir d'une interface qui permet de :

- passer les noms des fichiers,
- sélectionner les variables d'intérêt, les variables explicatives, les variables de calcul (poids, phases...),
- définir les paramètres d'exécution et sélectionner les traitements.

A) Préparation des fichiers

Il s'agit d'introduire dans les fichiers les identifiants qui permettront leur rapprochement.

Le fichier géographique (GEO)

Le fichier géographique apporte des informations auxiliaires comme les effectifs des unités sondées, ou leur taille, données nécessaires au calcul des probabilités d'inclusion. Ces données ne peuvent, en général, pas être générées à partir du fichier d'enquête, et proviennent donc de sources annexes (par exemple, le recensement de la population). Il faudra généralement harmoniser ses identifiants géographiques et ceux du fichier de données : mêmes types (numérique ou caractère sur n positions) et mêmes codes, comme l'exige le logiciel sous-jacent Sas.

Le fichier de données (DATA)

Pour pouvoir rapprocher le fichier de données du modèle représentant le sondage, il faut préciser à quel arc terminal de l'arbre des traitements se rapportent les données,

en mentionnant dans une variable supplémentaire (appelée *NINFFIC*) le code de la feuille concernée (code à 2 lettres). Ainsi, dans l'exemple de la page 5, le logement *log6* recevra le code feuille "AC" , le logement log11 recevra le code "AC".

On injecte ce code dans chaque observation de la table des données SAS, à partir des identifiants géographiques.

Pour les enquêtes en plusieurs phases, aux données recueillies sur le terrain, il faut adjoindre les données des échantillons des phases précédentes qui, bien que mises à zéro par le logiciel, sont indispensables dans les calculs.

L'utilisateur doit en outre fournir dans la base des données (DATA) :

- la probabilité d'inclusion de la 2ème phase, pour les enquêtes en 2 phases Poissonnien,
- la probabilité d'inclusion de la 3ème phase, pour les enquêtes en 3 phases,
- le poids final *Wip* de l'enquête, retenu pour la diffusion des résultats ; ce poids pourra (lorsqu'il y aura eu calage) différer du poids global élaboré par le logiciel à partir du plan de sondage.

C'est ce poids *Wip* qui est retenu pour l'évaluation des totaux « pondérés », sur lesquels sont centrés les intervalles de confiance produits par le logiciel.

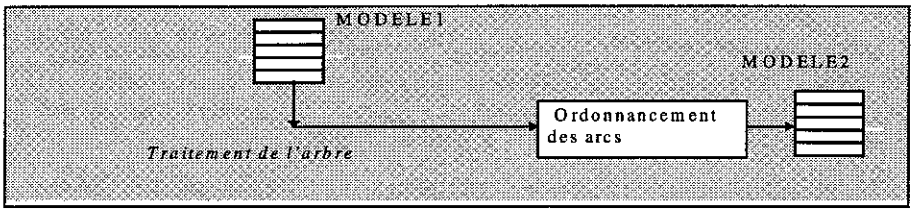
Le modèle du sondage (MODELE)

On crée la table Sas décrivant le modèle du sondage, chaque observation correspondant à un arc décrivant un tirage élémentaire, avec les données suivantes, au minimum :

- identifiants des extrémités de l'arc, sous la forme de codes à 2 lettres,
- type de tirage élémentaire : aléatoire simple, aléatoire simple équilibré, systématique, à probabilités d'inclusion inégales, exhaustif, total,
- noms des entités tirées : ex. REGION CANTON COMMUNE,
- noms des entités à l'intérieur desquelles on a réalisé le tirage : ex. REGION CANTON

B) Contrôle du modèle et génération

Une fois saisi le modèle, un module permet d'en vérifier certaines propriétés (connexité, absence de réseau...),

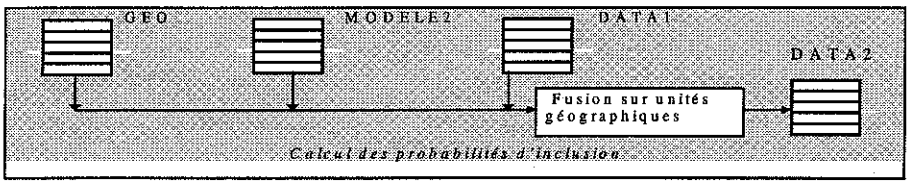


et de générer de nouvelles variables sur la topologie de l'arbre : la structure des formules impose de traiter les arcs dans un certain ordre pour suivre la règle suivante :

On peut traiter un arc lorsqu'il aboutit à une feuille ou lorsque tous ses arcs inférieurs ont été traités.

C) Calcul des probabilités d'inclusion

Dans cette étape, le logiciel calcule les probabilités d'inclusion élémentaires de la première phase, i.e. celles relatives à un sondage élémentaire, à partir :



- de données contenues dans le fichier de données : nombre d'entités tirées,
- de données issues du fichier "géographique" : taille des entités tirées, taille des entités d'agrégation,
- du type de sondage élémentaire, lorsque ces probabilités sont absentes du fichier de données.

Cas où les probabilités d'inclusion dépassent 1 : pour les sondages proportionnels à la taille, le calcul brut peut conduire à des valeurs supérieures à 1. Les données sont alors triées par ordre décroissant de taille des entités tirées ; on met à 1 la plus grande probabilité qui dépasse 1, et on recalcule les probabilités d'inclusion des autres entités tirées, après avoir mis à jour la taille de l'échantillon et la taille de l'ensemble des entités ; si à nouveau une valeur de probabilité dépasse 1, on réitère le processus.

Une fois toutes les probabilités d'inclusion élémentaires évaluées, leur produit fournit la probabilité globale d'inclusion de la première phase (égale à l'inverse du poids de sondage).

Pour les enquêtes en plusieurs phases dont la 2ème phase est stratifiée, le logiciel calcule la probabilité d'inclusion de la 2ème phase comme le rapport *nh/NH, nh* étant l'effectif pour la strate *h* dans l'échantillon 2ème phase, et *NH* l'effectif pour la strate *h* dans l'échantillon 1ère phase, ce qui requiert la présence des échantillons de la ou des phases précédentes dans la base des données de l'enquête.

Pour un sondage Poissonnien, le taux de sondage de la 2ème phase doit être présent dans le fichier de l'enquête (DATA).

La probabilité d'inclusion finale est égale au produit des probabilités d'inclusion relatives à chacune des phases.

D) Définition des variables d'intérêt

Les variables d'intérêt sont entrées sous la forme d'une liste de paramètres passée à une macro. La syntaxe du logiciel SAS est acceptée pour les variables de groupe, par exemple x1-x5 ou a--d ; cependant le nom d'une variable ne doit pas dépasser 7 caractères (pour une variable simple), ou 16 caractères (pour une variable de groupe) .

C'est dans cette étape que l'on définit les variables d'intérêt sur lesquelles on veut lancer le calcul des estimateurs, et que l'on prépare le fichier de données :

- en mettant à zéro les variables d'intérêt de la première phase (pour les enquêtes en 2 phases),

- en mettant à zéro les variables d'intérêt des 2 premières phases (pour les enquêtes en 3 phases).

E) Statistiques complexes : comment estimer des ratios ?

Les statistiques complexes (appelées fonctions dans ce document) sont traitées à l'aide d'un mécanisme particulier, qui permet de les décrire par appel de fonctions élémentaires (par exemple les fonctions arithmétiques somme, différence, produit, ratio ou exponentielles).

Base théorique de l'évaluation des estimateurs pour les fonctions

Etant donné un échantillon **s** d'une population **P**, il s'agit d'évaluer la précision d'une fonction construite sur des totaux : par exemple le revenu moyen par personne, à partir du revenu et du nombre de personnes du ménage.

Le logiciel procède par linéarisation des fonctions à estimer, à l'aide de la formule de Taylor, de développement en série à partir des dérivées partielles, approche formalisée par Woodruff en 1971, (démarche suivie par l'institut Statistics Sweden dans le logiciel CLAN). Il reprend également les principes de programmation du logiciel CLAN, à partir de fonctions élémentaires écrites en macro SAS et de fonctions "utilisateurs" bâties sur ces fonctions dérivables.

Ainsi, on remplace une fonction de totaux, par le total d'une fonction définie sur chaque observation, fonction linéaire des variables observées ; on sait alors estimer la variance de ces totaux de fonctions linéaires.

Cette démarche est pertinente pour les échantillons suffisamment représentatifs pour que l'on puisse négliger dans le développement en série de Taylor, les termes de rang supérieur ou égal à 2.

Soit, par exemple, à évaluer la fonction : $\theta = \frac{t1}{t2}$ (revenu moyen par ménage), avec :

* y1 représentant le revenu du ménage, t1 = Σy1,

* y2 représentant le nombre de ménages, t2 = Σy2.

On pose : $z_k = \frac{\partial \theta}{\partial t1} * y_{1k} + \frac{\partial \theta}{\partial t2} * y_{2k}$

Alors : $z_k = \frac{1}{\hat{t}_2} * y_{1k} - \frac{\hat{t}_1}{\hat{t}^2_2} * y_{2k}$

On démontre que la variance du total de z_k sur la population est approximativement égale à la variance de θ, c'est-à-dire à l'erreur quadratique moyenne, si l'on néglige le biais.

Ainsi on linéarise les fonctions par la formule de Taylor :

$$\hat{\theta} - \theta = \sum_{j=1}^{J} f_j^{'}(t)(\hat{t}_j - t_j) \text{ où } f_j'(t) = \frac{\partial f(t)}{\partial t_j}$$

et la formule de transformation de Woodruff :

$$z_k = \sum_{j=1}^{J} f_j'(\hat{t})\, y_{jk}$$

puis, à l'aide de fonctions de base (+,-,*,/), on génère les calculs d'estimateurs à effectuer. Ainsi, pour toute fonction rationnelle que l'on peut écrire à partir des opérateurs de base, on peut exprimer la dérivée à partir :

- des règles de base pour la dérivation de fonctions de fonctions,

- de la dérivation de quatre fonctions arithmétiques de base (addition, soustraction, multiplication, division) ou d'autres fonctions (exponentielle).

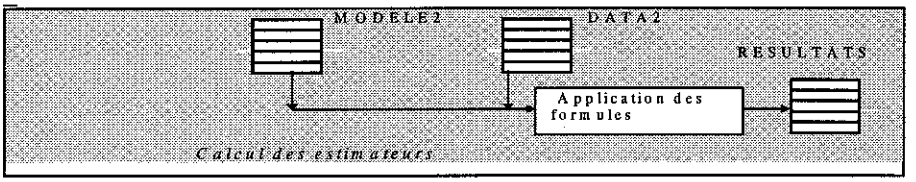
Définition d'une statistique complexe par l'utilisateur

Il reste à la charge de l'utilisateur la définition de la statistique complexe, à l'aide d'instructions qui se présentent ainsi :

%DIV(RATIO,REVENU,POP), si l'on cherche à mesurer le revenu moyen (RATIO), à partir de la variable REVENU et de l'indicatrice POP présentes dans le fichier : cette instruction permet de créer la variable z_k , dont la variance donnera celle de l'estimateur de la variable RATIO = REVENU/POP .

F) Calcul des estimateurs

Des modules spécifiques déroulent les formules de calcul pour obtenir :



- l'estimateur du total de la variable ou l'estimateur du total de la statistique complexe, à partir des probabilités d'inclusion du logiciel, qui ne tiennent pas compte d'éventuels calages,
- l'estimateur de la variance de cet estimateur,
- l'estimateur du total pondéré de la variable ou l'estimateur du total de la statistique complexe, à partir du poids final W_{ip},
- l'intervalle de confiance centré sur le total pondéré,
- l'effet de sondage ("design effect") sur option.

Ce calcul est fondé sur des formules propres au type de sondage, sur les variables d'intérêt présentes dans le fichier de données, sur les probabilités d'inclusion calculées en partie ou en totalité par le logiciel.

Le logiciel distingue 2 classes de formules :

- celles qui sont conduites au niveau des arcs du modèle, en suivant l'ordre défini par la règle du §2.b (ordre déterminé au cours de la génération du modèle), et qui font intervenir les probabilités d'inclusion élémentaires,
- celles qui sont appliquées sur l'ensemble des données, car elles ne font référence qu'aux probabilités d'inclusion globales.

Pour les enquêtes stratifiées, les formules exigent la création de variables intermédiaires au niveau de chacune des strates, pour chaque variable d'intérêt (ou statistique complexe), autant de variables 'strates' qu'il y a de strates, plus une.

Les valeurs manquantes sur les variables d'intérêt n'arrêtent pas les calculs, mais altèrent les résultats ; le logiciel en donne la fréquence.

On peut aussi faire appel à une méthode simplifiée de calcul des variances à partir des poids et d'un modèle dérivé du modèle théorique du plan de sondage.

Les formules de calcul aboutissent à des résultats erronés ou incomplets, dans les cas suivants:

- données manquantes,
- population de l'échantillon égale à 1.

Dans ces cas là, le logiciel met en oeuvre les traitements par défaut de Sas sur les données manquantes : elles sont ignorées.

4. Domaine d'application

Le logiciel couvre le domaine suivant :

- les sondages à 1 ou n degrés en une phase,
- les sondages à 1 ou n degrés, en deux phases, lorsque la seconde phase est un sondage aléatoire simple stratifié,
- les sondages à 1 ou n degrés, en deux phases, lorsque la seconde phase est un sondage Poissonnien,
- les sondages à 1 ou n degrés, en trois phases, lorsque la deuxième phase est un sondage aléatoire simple stratifié, et la troisième phase un sondage Poissonnien : ceci permet de traiter les enquêtes en deux phases stratifiées avec une correction de non réponse, en considérant cette dernière comme un sondage Poissonnien.

Pour les sondages élémentaires, les formules programmées actuellement portent sur les types suivants :

- sondage aléatoire simple sans remise,
- sondage à probabilités inégales (en particulier à probabilités proportionnelles à la taille),
- sondage systématique à probabilités égales,
- sondage stratifié,
- sondage équilibré.

La statistique d'intérêt peut être le total d'une variable, ou une statistique complexe, fonction de plusieurs totaux de variables. Dans ce dernier cas, la méthode de linéarisation, programmée dans le logiciel, permet de se ramener à l'estimation de la variance du total d'une variable synthétique.

Dans toutes ces configurations, le logiciel offre la possibilité de calculer les estimateurs d'Horvitz-Thompson et leur variance, et l'effet de sondage (Design Effect).

Pour les enquêtes dont les données ont été redressées par le logiciel Calmar, la précision est donnée par la variance des estimateurs sur les résidus, (et non plus sur les variables), calculés à partir d'une régression (procédure GML de SAS) sur les variables explicatives (numériques et caractères) passées au logiciel.

5. Mise en œuvre du logiciel

Les programmes sont écrits en langage Macro de SAS, avec appel de procédures courantes, et sans utilisation d'outil particulier de ce logiciel statistique en exécution. Ils tournent sous les systèmes d'exploitation MVS (IBM) et Windows (Microsoft).

La version actuelle est issue de spécifications d'études, et non de production. Elle dispose d'une interface interactive (en Sas Scl, ergonomie MVS), qui permet d'entrer les principaux paramètres, et de générer les macros d'exécution.

Principaux paramètres décrivant l'enquête

- nombre de phases de l'enquête : {1,2,3}, et nom de la variable phase,
- méthode de calcul : {simple, stratifié, Poissonnien, Ultimate Clusters},
- poids final,
- probabilités de réponse pour la 2ème phase Poissonnien, et la 3ème phase,
- liste de variables définissant les strates, pour les phases stratifiées,
- listes des variables explicatives pour les enquêtes ayant été redressées par le logiciel Calmar.

Les autres paramètres portent sur les noms des fichiers en entrée, la liste des variables d'intérêt et les options d'édition.

Ressources nécessaires

- espace disque : environ 6 fois la taille du fichier d'enquête (pour les fichiers intermédiaires, pour la création des fonctions, des variables strates, des résidus...),
- logiciel Sas version sur micro 6.11 sous Windows et 6.08 sur MVS.

Délais d'apprentissage et d'exécution du logiciel

L'opération de modélisation du sondage, lorsqu'il s'agit d'un sondage complexe, est plus longue que la mise en oeuvre du logiciel. En effet, le processus de modélisation, malgré son apparente simplicité, demande pour sa compréhension un minimum d'apprentissage.

Il faut compter environ une demi-journée pour se familiariser avec le lancement du logiciel ; en revanche les tâches préliminaires pourront demander davantage de temps, pour :

- déterminer avec rigueur les caractéristiques du sondage à l'origine de l'enquête,
- rassembler les sources contenant les effectifs nécessaires au calcul des probabilités d'inclusion,
- harmoniser les identifiants des diverses sources.

Pour les enquêtes de l'Institut issues du dernier échantillon-maître, l'essentiel de ce travail préparatoire a déjà été accompli, il ne reste généralement qu'à adapter à la marge le modèle, au niveau des derniers tirages.

Bibliographie

DEVILLE J.C. Estimation de précision de données d'enquête, Insee F9211, 1992 ;
Calcul de l'effet de sondage, Mars 96.

ISNARD M. Validation expérimentale du modèle théorique, Décembre 1992.

NEROS B. Base de sondage des logements neufs, 589/F010, Octobre 1993.

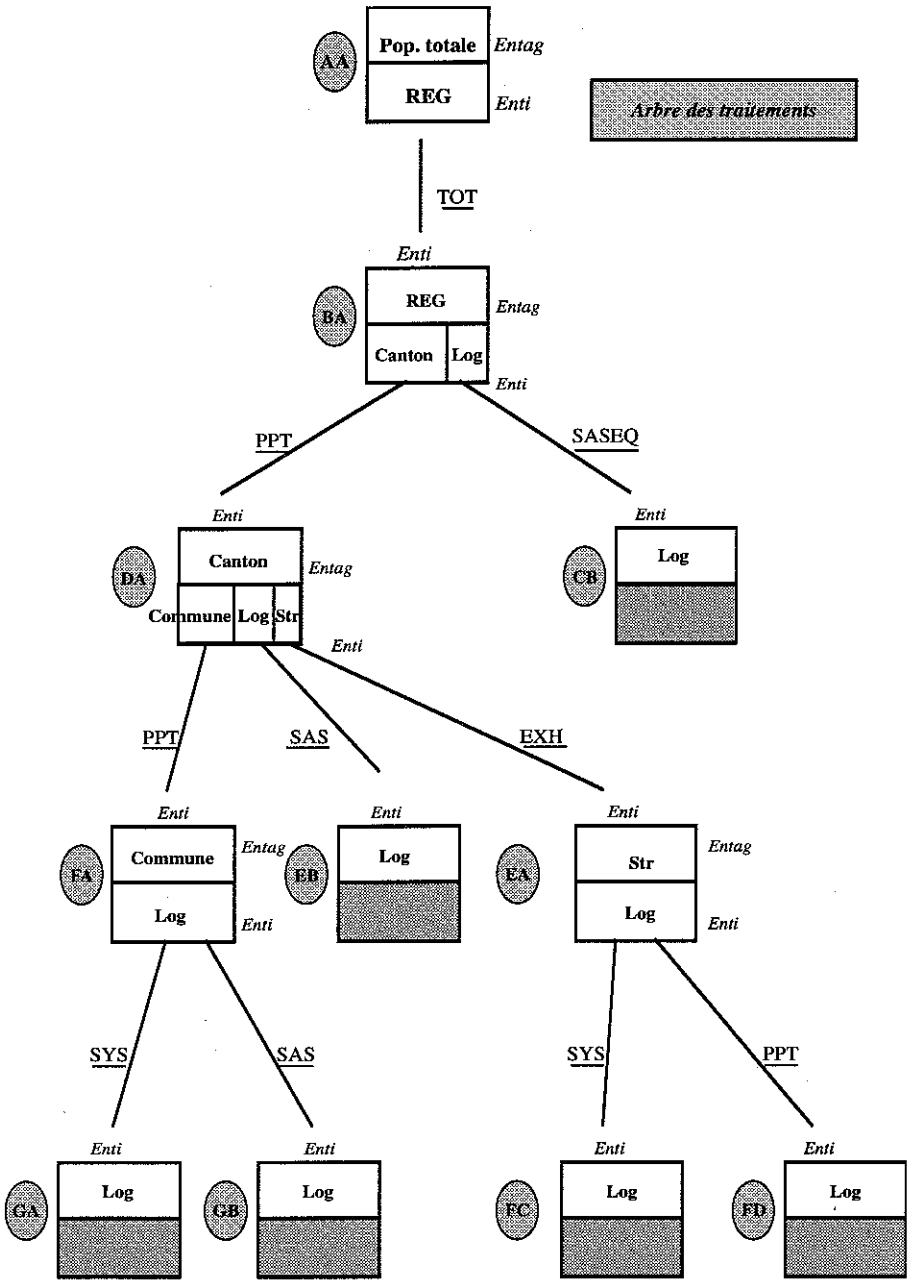
Projet sur l'estimation de précision :
fichier géographique, et fichier de données issu de l'enquête, 688/F010, Mars 1994.

SAUTORY O. Estimation de la variance dans un plan de sondage à plusieurs degrés, Insee, 3 Nov 92.

DUPONT F. Non réponse et sondages en plusieurs phases dans le logiciel de calcul de précision, Insee 4/F420, 10 Juin 94.

CARON N. Sondage 2ème phase Poissonnien dans le logiciel Poulpe
Insee 958/F410, 29 Jan 96.
Calcul simplifié de la variance, Insee 968/F410, 22 Février 96.

Modèle de sondage à plusieurs degrés



ANNEXE 2

Estimation de précision des enquêtes : schéma général