CALCUL DE PRÉCISION AU MOYEN DU LOGICIEL POULPE DANS L'ENQUÊTE HID

Pascal ARDILLY (*), Odile JOINVILLE (**), Pierre MORMICHE (***)

(*) Insee, Unité Méthodes Statistiques
(**) Institut de Statistique de l'Université de Paris
(***) Insee, Département de la Démographie

Introduction

La recherche d'estimateurs de variance sur l'enquête HID réalisée en population générale fin 1999 se heurte à des difficultés majeures. Pour dire les choses rapidement, (1) le plan de sondage est très compliqué ; (2) la première étape de l'opération porte sur des échantillons de taille tout à fait inhabituelle pour des enquêtes auprès des ménages, ce qui alourdit de façon considérable l'exécution des macros d'estimation ; (3) malgré cela on souhaite pouvoir fournir un outillage simple aux équipes de chercheurs amenées à exploiter l'enquête.

Même si tout n'est pas réglé à cet instant, on a pu cependant vérifier que Poulpe présente un éventail suffisant d'options pour prendre en compte cette complexité, et livre non seulement des estimations de variance mais aussi des estimations sur les effets de sondage. Il permet aussi d'apprécier les effets de redressements sur la précision.

Dans ce qui suit, on présentera d'abord les principaux aspects de la gymnastique à laquelle le modèle - et ses utilisateurs - ont dû se soumettre pour faire rentrer le mécanisme dans le moule ; on présentera ensuite les corrections auxquelles on a dû (et pu) procéder et l'orientation du travail à venir.

1. Faire entrer HID dans le « moule » Poulpe

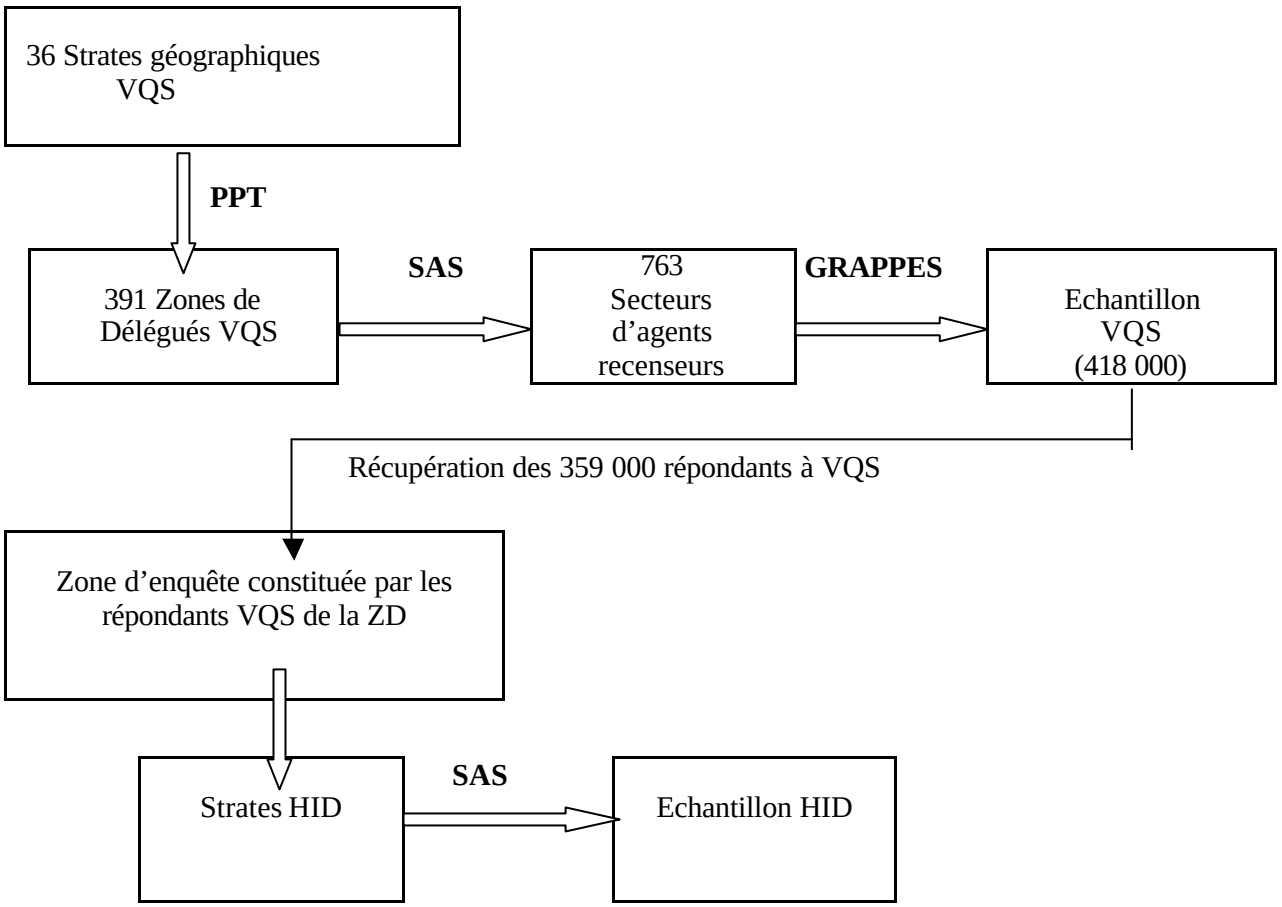
1.1 Le plan de sondage de l'enquête

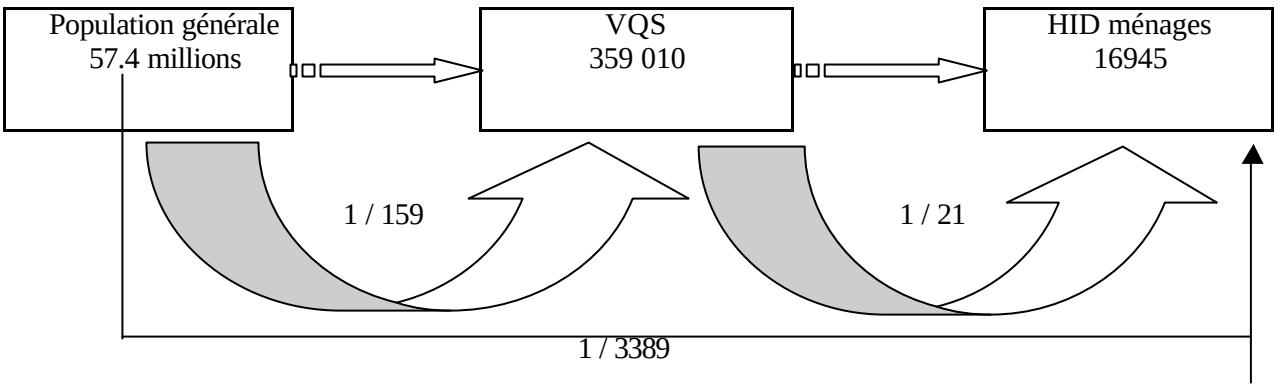
Le plan de sondage de l'enquête est un plan en deux phases : une première phase correspond à la « pré enquête » Vie Quotidienne et Santé (VQS), associée au recensement de mars 1999 et réalisée auprès d'un échantillon de 418 000 personnes, dont 359 000 réponses exploitables. L'échantillon de la seconde phase (21 970 personnes) a été tiré parmi ces 359 000 répondants, en fonction notamment des réponses recueillies à VQS ; 16 945 personnes ont répondu à cette seconde partie, dite HID (Handicaps-Incapacités-Dépendance).

1.1.1 Le plan de sondage de la pré enquête VQS

On peut distinguer quatre étapes :

- en premier, le territoire métropolitain a été réparti en 36 strates géographiques. On commence par stratifier le territoire national en régions administratives. Certaines régions sont scindées de nouveau, afin d'y distinguer un ou plusieurs département(s), voire des fractions de département. En effet, sept conseils généraux ont demandé des extensions départementales d'échantillon VQS (ainsi que la région Haute-Normandie) et cinq départements ont demandé des extensions locales d'échantillon pour l'enquête histoire familiale (EHF) – cette dernière étant étroitement liée à VQS. On obtient in fine 36 strates.





- en second, on a échantillonné dans chaque strate, à titre d'unités primaires, des zones de délégué (ZD), qui sont des ensembles connexes d'environ 15 000 habitants en moyenne – dont l'intitulé rappelle qu'il s'agit des zones de compétence des délégués du Recensement général de la population de mars 1999. Le tirage de ces zones (3553 ZD ont été découpées sur l'ensemble du territoire) a été effectué à probabilités proportionnelles à leur taille en nombre de personnes recensées en 1990. Dans les strates à extension VQS ou EHF, le taux de sondage des ZD est plus fort.. Au total, 425 zones de délégué on été tirées, dont 391 utilisées pour la collecte HID.

- dans chaque ZD tirée, on a échantillonné de zéro à six secteurs d'agent recenseur (SAR). Un secteur d'agent recenseur est un ensemble connexe de logements dans lequel habitent en moyenne 500 à 600 personnes, constitué soit de communes s'il s'agit de petites communes, soit de districts si on se situe dans une « grande » commune. Les SAR sont tirés par sondage aléatoire simple. Le nombre de ces unités secondaires tirées est assez variable selon la ZD ; en fait, il dépend essentiellement de la nature de la strate dans laquelle on se situe : dans une zone sans extension VQS, on tire le plus souvent un ou deux SAR par ZD, dans une zone à extension VQS on augmente sensiblement la taille de l'échantillon mais dans une zone à extension EHF on a tendance au contraire à la diminuer. Au total 763 SAR ont été tirés au sein des 391 ZD.

- enfin toute la population des secteurs désignés (416 000 personnes) a fait partie de l'échantillon - il s'agit donc d'un sondage en grappe si on fait abstraction de la non réponse - , mais seule une partie (359 000) a répondu. On a assimilé sa sélection à un tirage aléatoire simple dans chaque secteur.

1.1.2 Le plan de sondage de l'enquête HID

L'échantillon VQS est un échantillon d'individus – et non de ménages – qui a été post-stratifié en croisant trois dimensions : une dimension de niveau de handicap, une dimension d'age et une dimension géographique.

On a d'abord distingué, à partir d'une synthèse des réponses fournies au questionnaire de la pré enquête, 6 groupes définissant des handicaps croissants :

Groupe 1 (78.6 %) : personnes déclarant ne souffrir d'aucune difficulté ;

Groupe 2 (6.6 %) : personnes déclarant une seule difficulté ;

Groupe 3 (4.4 %) : personne déclarant « avoir un handicap » ou « avoir demandé une reconnaissance » ou souffrir d'une « limitation d'activité » ou dépendre d'une aide humaine ou souffrir de plusieurs autres difficultés ;

Groupe 4 (2.6 %) : personne déclarant « avoir un handicap » ou « avoir demandé une reconnaissance » et personnes déclarant souffrir d'une «limitation d'activité », déclaration appuyée par des items d'aide humaine ou technique ou plusieurs autres ;

Groupe 5 (3.8 %) : personne déclarant « avoir un handicap » ou « avoir demandé une reconnaissance » , déclaration fortement appuyée par d'autres items ;

Groupe 6 (4.0 %) : personne déclarant avoir obtenu une reconnaissance de son handicap, plus les enfants et les adolescents de moins de 16 ans inscrits dans une classe ou un établissement spécialisés.

Ensuite on a distingué, au sein des groupes 1, 2, 5 et 6, les personnes de moins de 70 ans et celles de 70 ans et plus. Globalement, on a donc construit 10 post strates croisant le degré de handicap et l'age :

Post strate 1 : groupe VQS 1 et age < 70 ans
Post strate 2 : groupe VQS 1 et age >= 70 ans
Post strate 3 : groupe VQS 2 et age < 70 ans
Post strate 4 : groupe VQS 2 et age >= 70 ans
Post strate 5 : groupe VQS 3 , tout age
Post strate 6 : groupe VQS 4 , tout age
Post strate 7 : groupe VQS 5 et age < 70 ans
Post strate 8 : groupe VQS 5 et age >= 70 ans
Post strate 9 : groupe VQS 6 et age < 70 ans
Post strate 10 : groupe VQS 6 et age >= 70 ans

Par ailleurs, on a introduit la notion de «zone d'enquête » pour constituer un autre critère de post stratification : la zone d'enquête ne peut se définir simplement et de manière universelle, car c'est un découpage ad hoc effectué en fonction de la configuration du terrain. Le plus souvent, la zone d'enquête coïncide avec la ZD, mais on trouve d'autres découpages. Au total, il y a un peu plus de 400 zones d'enquête.

La post stratification complète a croisé les 10 post strates définies ci-dessus avec l'ensemble des zones d'enquête. Les taux de deuxième phase dépendent de la zone d'enquête, mais ils sont établis sous contrainte de proportionnalité au vecteur fixe suivant :

	Coefficients de tirage proportionnels à :
groupe 01	0,68
groupe 02	18,75
groupe 03	5,62
groupe 04	19,44
groupe 05	20
groupe 06	30
groupe 07	52
groupe 08	32
groupe 09	65
groupe 10	40

Il s'agit donc d'un tirage à probabilités très fortement inégales. L'éventail des probabilités d'inclusion est encore élargi par les contraintes liées aux extensions locales : d'une part dans sept des huit zones (six départements et une région) où l'échantillon VQS a été surdimensionné, les taux de tirage de HID dans VQS ont dû en contrepartie être diminués ; d'autre part, dans l'Hérault, où en plus de l'extension VQS, a été réalisée une extension de HID elle-même, les taux de tirage HID/VQS ont dû à l'inverse être accrus.

Le passage des 359 010 répondants VQS aux 16 945 répondants HID s'est effectué en deux étapes :

- le tirage post-stratifié décrit ci-dessus a sélectionné 21 970 personnes ;
- la phase de collecte, qui a comporté des échecs, a abouti à un échantillon de 16 945 répondants. La perte est essentiellement due aux refus, aux perdus de vue entre la date du recensement et la date de l'enquête (on recherche un individu donné, mais on ne parvient pas toujours à le suivre sur son nouveau lieu de résidence lorsqu'il déménage), aux absents de longue durée, aux décès et aux entrées en institution sanitaire et sociale depuis la date du recensement. Cela représente un taux de déchet de 22.2 %. La prise en compte de cette non réponse constitue la troisième phase de tirage.

1.2 Principe de fonctionnement du logiciel Poulpe

Ce logiciel de calcul de variance a été mis au point à l'Unité de méthodologie statistique de l'INSEE. Il permet d'effectuer des estimations de variance dans le cadre de plans complexes à plusieurs degrés avec des tirages à probabilités inégales, y compris pour des enquêtes en deux phases. Il peut estimer la variance liée à la non réponse par une modélisation assimilant la non réponse à un tirage de Poisson, ce qui correspond techniquement à une troisième phase de tirage.

L'estimation de variance dans les plans complexes combine différentes expressions, chacune fournissant une estimation relative à une étape « élémentaire » du plan : ainsi, les traitements de base distingués en tant que briques élémentaires de l'architecture d'ensemble concernent successivement les tirages à probabilités inégales, à plusieurs degrés, le tirage en deux phases avec post-stratification et le tirage de Poisson pour modéliser le processus de non réponse.

Le tirage à probabilités inégales pose un sérieux problème parce que la variance qui lui correspond fait intervenir des probabilités de sélection double. Or ces dernières sont généralement incalculables, et très difficiles à approximer. En pratique, on utilise des approximations de l'estimateur de variance dont la complexité reste «raisonnable ». Dans Poulpe, on a implanté une expression proposée par Jean-Claude Deville (Crest, Laboratoire de statistique d'enquêtes), qui procède du principe suivant : dans la classe des plans de taille fixe n, le plan de sondage d'entropie maximale est le plan de Poisson contraint, c'est-à-dire le plan de Poisson conditionné par la taille fixée de l'échantillon. Il se trouve que l'on parvient assez facilement à calculer la variance d'un tel plan, sous quelques hypothèses techniques tout à fait réalistes. On trouve l'expression suivante, s désignant l'échantillon tiré :

$$\hat{V} = \frac{n}{n-1} \sum_{k \in s} (1 - \Pi_k) \left(\frac{y_k}{\Pi_k} - \sum_{k \in s} a_k \frac{y_k}{\Pi_k} \right)^2$$

où $$a_k = \frac{1 - \Pi_k}{\sum_{k \in s} (1 - \Pi_k)}$$

et Π_k est la probabilité de sélection de l'unité k.

Si on considère que l'algorithme de sélection est à entropie très forte, comme ce doit être le cas par exemple avec un tirage systématique dont l'ordre du fichier est aléatoirisé, alors la formule précédente doit être une bonne approximation de la réalité.

Pris isolément, le tirage à deux degrés est un peu plus simple à traiter sur le plan théorique : par un conditionnement adéquat, on parvient à exprimer sa variance sous forme d'une somme de deux termes : le premier représente la part de variance « inter » correspondant à l'incertitude générée par le tirage des unités primaires, le second traduit la variance «intra » liée au tirage aléatoire des unités secondaires dans l'unité primaire.

La difficulté avec les tirages à plusieurs degrés provient plutôt de l'implémentation des formules dans les programmes informatiques lorsqu'il y a plusieurs degrés successifs, surtout avec un nombre élevé de strates et lorsqu'on pratique à chaque degré des tirages complexes (comme par exemple des tirages à probabilités inégales) : dans ces conditions, les expressions mathématiques deviennent très vite inextricables et ne peuvent pas être programmées en une seule fois. Une des forces de Poulpe est son caractère « universel » car il s'appuie sur une programmation récursive qui évite de se noyer dans des expressions analytiques épouvantablement compliquées. Pour cela, il utilise l'expression suivante donnée par Raj :

$$\hat{V} = f\left(\hat{Y}_i \middle| i \in U\right) + \sum_{i \in U} w_{is} \hat{V}_i$$

où :

- $f\left(Y_i \middle| i \in U\right) = \hat{V}_1\left(\sum_{i \in U} w_{is}\ Y_i\right)$, variance estimée du premier degré de sondage dans le cas où les vrais totaux par UP, notés Y_i, sont connus (dans l'expression ci-dessus, on remplace donc les vrais totaux par leurs estimateurs sans biais).
- $\hat{V}_i = \hat{V}_{2/1}\left(\hat{Y}_i\right)$, estimateur sans biais de variance au second degré de sondage, dans l'$UP\ i$.
- w_{is} est le poids de l'$UP\ i$, tel que $\sum_{i \in U} w_{is}\hat{Y}_i$ estime le vrai total Y sans biais.

Le tirage en deux phases conduit à une expression de variance qui comprend deux termes : un premier terme lié au tirage de première phase - donc à l'échantillonnage VQS - qu'on ne reproduira pas ici car il s'agit de la résultante du plan complexe de première phase et un second terme dû à l'échantillonnage au sein de chacune des post-strates. Le second terme varie en $1/n_{h,2}$ où $n_{h,2}$ représente la taille de l'échantillon tiré en post-strate h et vaut, en notant $n_{h,1}$ la taille de l'échantillon première phase recoupant la post-strate h :

$$\hat{V} = \sum_{h=1}^{H} \frac{1 - \dfrac{n_{h,2}}{n_{h,1}}}{n_{h,2}} \cdot n_{h,1}^2 \cdot \frac{1}{n_{h,2} - 1} \sum_{k \in S_{h,2}} \left(z_k - \bar{z}_h\right)^2$$

où $z_k = y_k / \Pi_k$

et $\bar{z}_h = \dfrac{1}{n_{h,2}} \sum_{k \in S_{h,2}} z_k$

Enfin, la variance générée par la non réponse est obtenue à partir d'une modélisation par un schéma de Poisson : on considère que chaque individu échantillonné répond avec une probabilité donnée q_k ou ne répond pas avec la probabilité complémentaire $\left(1 - \theta_k\right)$. Le processus de réponse est indépendant d'un individu à l'autre. La variance associée à un processus de Poisson à probabilités inégales q_k conduisant à un échantillon S est :

$$\hat{V} = \sum_{k \in s} \frac{1 - \theta_k}{\theta_k} \cdot y_k^2$$

La prise en compte de l'intégralité du plan de sondage conduit à la constitution de deux fichiers fondamentaux. Un premier fichier, dit «arbre descriptif du plan de sondage », décrit la nature des unités de sondage sollicitées à chaque degré et le mode d'échantillonnage retenu : on précise tous ces paramètres du tirage afin que la formule de Raj puisse être appliquée en combinaison avec l'expression de variance à probabilité inégales, en tenant compte des étapes préalables de stratification. Un second fichier, dit « fichier géographique », donne la liste des unités d'échantillonnage sollicitées à chaque degré de sondage, avec les informations nécessaires pour le calcul de leurs probabilités de sélection.

Lorsqu'il s'agit de calculer des variances d'estimateurs complexes (donc non linéaires), le travail préparatoire se partage entre le logiciel et l'utilisateur : le premier effectue des calculs à partir de fonctions de base telles que somme, ratio, produit ou exponentielle, et le second a la charge de la linéarisation de l'estimateur qui aboutit à une décomposition selon les fonctions de base en question.

A noter une difficulté - sérieuse dans notre cas - au niveau du traitement des tailles d'échantillon égales à 1 : le logiciel, lorsqu'il rencontre ce cas de figure, ignore purement et simplement la variance associée au tirage. Autrement dit, il fait comme si la variance était nulle (alors que, justement, elle est grande puisque la taille de l'échantillon est extrêmement petite !).

Le logiciel fournit également des estimations d'effets de sondage (« design effect »), définis comme les rapports de la variance obtenue avec le plan de sondage sur la variance que l'on obtiendrait si on pratiquait un sondage aléatoire simple de même taille. Il permet aussi de prendre en compte l'effet d'éventuels redressements en injectant dans les formules précédentes les résidus des régressions des variables d'intérêt sur les variables de calage.

On récupère, pour les variables d'intérêt dont on cherche la précision, l'estimation du total obtenue avec les poids présents dans le fichier d'origine (poids bruts ou redressés, au choix), l'estimation du même total résultant d'une pondération globale définie en multipliant les poids calculés « localement » par le logiciel à chaque degré de sondage (donc en effectuant un produit des probabilités d'inclusion à chaque étape), une estimation de l'écart type et une estimation des bornes de l'intervalle de confiance à 95% pour ce total. On peut en déduire immédiatement le coefficient de variation, défini comme le rapport de écart type estimé au total estimé.

Concrètement, Poulpe se présente sous forme d'une succession d'écrans de dialogue, au travers desquels on entre des paramètres, utilisés ensuite par des programmes SAS.
On distingue 5 étapes fondamentales à activer successivement pour obtenir des résultats :

- enrichissement et contrôle de l'arbre décrivant le plan de sondage (ARBGEN)
- calcul des probabilités d'inclusion au niveau de chaque étape composant le plan de sondage (CALPII)
- chargement de la liste des variables d'intérêt (CHARLIS)
- calcul proprement dit des estimations de variance des totaux des variables d'intérêt précédemment sélectionnées (ESTIVAR)
- calcul éventuel d'estimation de variance pour des statistiques complexes (notamment des proportions de personnes concernées dans diverses sous-populations) - utilisant la méthode de linéarisation (ESTIFON)

La convivialité du logiciel est actuellement perfectible, et il ne peut s'utiliser qu'après une phase de préparation de fichiers qui peut être lourde si le plan est complexe.
Le lecteur intéressé par le fonctionnement de Poulpe pourra consulter la documentation technique du logiciel, qui est très complète.

1.3 La description du plan de sondage dans Poulpe

1.3.1 Modélisation informatique théorique du plan de sondage

Le plan décrit à Poulpe est un plan en trois phases.

En première phase, l'échantillon est stratifié à 3 degrés : un code de strate (36 modalités) figure dans le fichier géographique, et on y précise les zones de délégué tirées (unités primaires) et les secteurs d'agents recenseurs tirés (unités secondaires), chacun étant repéré par un identifiant spécifique. On fournit les tailles des zones de délégués (nombre de personnes recensées en 1990) et le nombre de secteurs d'agent recenseur dans chaque zone de délégué tirée, afin que le logiciel puisse calculer les probabilités de sélection de chacune de ces unités. Le troisième degré correspond à la sélection des répondants VQS : on a retenu un sondage aléatoire simple pour modéliser la réponse, par opposition à un tirage de Poisson qui n'est pas une modélisation permise par Poulpe lorsqu'on ne se situe pas en dernière étape du processus.

En seconde phase, il s'agit d'un tirage post-stratifié avec sondage aléatoire simple (de taille fixe) dans chaque post-strate. La variable repérant les post strates est une variable qualitative présente dans le fichier des données individuelles. A ce niveau, il est nécessaire de conserver l'intégralité de l'échantillon première phase, ici l'échantillon VQS, afin que les calculs de taux de sondage de seconde phase puissent être effectués par le logiciel.

La troisième phase renvoie au tirage de Poisson : la probabilité de réponse à HID (conditionnée au fait que l'on répond à VQS), estimée à partir d'un modèle logistique, figure au niveau de chaque individu répondant dans le fichier des données individuelles.

1.3.2 De la théorie à la pratique : quelques difficultés

A titre préliminaire, nous avons fait abstraction des modifications intervenues sur les échantillons originels, c'est-à-dire résultant directement du tirage aléatoire : Pour des raisons de terrain, il est arrivé en effet, à plusieurs reprises, que l'on remplace telle ou telle unité par une sélection raisonnée.

Les trois principales observations concernent le tirage des secteurs d'agents recenseurs, le tirage post-stratifié de l'échantillon HID parmi les répondants VQS et la phase de non-réponse lors de la collecte HID.

1. Le codage des secteurs d'agents recenseurs a été négligé dans les procédures du recensement, et a disparu de l'ensemble des fichiers de données. Aussi a-t-on choisi, dans un premier temps, de décrire l'étape de tirage des SAR comme si on avait en lieu et place procédé à un sondage aléatoire simple des districts composant ces SAR parmi l'ensemble des districts composant la zone de délégué « mère ».

   Il s'agit d'unités trois fois plus petites en moyenne (on compte 2 273 districts parmi les 763 SAR et 391 ZD retenues), mais de tailles très inégales (pour une taille moyenne de 184 personnes, la taille des districts VQS varie entre 2 et 1190). Cette forte dispersion des tailles provoque une augmentation de la variance, alors que l'augmentation du nombre d'unités tirées (3 fois plus nombreuses en moyenne) provoque une diminution de celle-ci. On peut raisonnablement penser que, globalement, cette simplification surestime in fine la vraie variance, car la dispersion des tailles des districts est un facteur très aggravant.

2. La description de la procédure de tirage HID dans VQS s'est heurtée à une limite technique du logiciel. Il s'agit d'un sondage post-stratifié à probabilités inégales selon la strate. Ces probabilités sont réparties en 3 660 strates (366 zones d'enquêtes à l'intérieur desquelles on tire dans chacune des dix strates HID selon des probabilités proportionnelles à l'échelle indiquée au chapitre 1.2.2). Or le logiciel n'admet actuellement en deuxième phase qu'un maximum de 99 post-strates.

   On a donc décidé dans un premier temps de distinguer seulement deux grandes zones : l'Hérault, où l'enquête HID a comporté une extension d'échantillon, et le reste de la métropole.

3. Pour ce qui concerne les non-réponses de la collecte HID, Poulpe permet de considérer cette étape comme une troisième phase autonome, où l'on procède à un tirage de type Poissonnien, selon des probabilités individuelles qui doivent être calculées par l'utilisateur et renseignées dans le fichier de données. Les probabilités ont été évaluées par modélisation logistique ; le modèle retenu, apparemment le plus performant, utilise 5 caractéristiques : le type de logement, la taille du ménage, la tranche d'unité urbaine, la « strate HID » et la tranche d'âge décennale.

Deux observations complémentaires :

- à plusieurs reprises (une fois pour les zones de délégués, dont une seule a été tirée dans la strate Corse, et 69 fois pour les districts) le plan de sondage décrit avait procédé au tirage d'une seule unité (ZD ou district). Dans ce cas, on sait qu'il est impossible de calculer une dispersion et on ne dispose donc pas d'un estimateur sans biais de variance. On a donc procédé à des regroupements d'unités « mères » - dits aussi « collapses » - (par exemple, on a regroupé la Corse avec la strate comprenant tous les départements de la Provence-Alpes-Côte d'Azur autres que les Bouches-du-Rhône) et considéré que le tirage avait été effectué dans l'unité regroupée - ce qui amène généralement à une surestimation de la variance lorsque les unités mères (les ZD) regroupées ne sont pas de tailles trop différentes.

- certaines unités n'ont pas été utilisées dans la suite du tirage. Par exemple, la strate 13 (petit morceau du département du Nord comportant une extension « langues » pour l'enquête associée sur l'Etude de l'Histoire Familiale, dans la zone flamingante) n'a pas comporté de zone de délégué VQS. Par exemple encore, certaines zones de délégué VQS n'ont pas été retenues - à la demande des services d'enquête de l'INSEE - pour le tirage de l'échantillon HID. Ces diverses « zones blanches » correspondent à une taille d'échantillon zéro et s'interprètent comme de la non réponse d'UP entières. Elles provoquent une sous-estimation (qui s'avère modeste) par le logiciel (qui utilise les inverses de probabilités de sondage) de la taille de la population métropolitaine globale et des totaux des variables d'intérêt. Il s'agit davantage d'un problème de biais que de variance, mais tout cela contribue néanmoins à accroître l'erreur globale

1.4 Résultats et fonctionnement de l'application

1.4.1 La taille du fichier de données

L'utilisation de Poulpe sur une enquête de la taille de VQS (le même problème surviendrait si on travaillait sur le fichier des individus de l'enquête Emploi, ou sur celui des individus de l'enquête sur l'Etude de l'Histoire Familiale) a permis de noter une anomalie imprévue : impossible de «faire tourner » la dernière étape du logiciel, celle qui procède à l'estimation des variances !

En analysant la Log, on soupçonna rapidement un problème de taille : un message ésotérique comportant un « E4 » apparaissait alors que la taille du fichier atteignait et dépassait précisément 100 000, i.e. 10^5. Or, en faisant tourner le logiciel sur le quart du fichier (82 000 observations), plus de problème ! Le responsable informatique alerté corrigeait cette limitation en moins de 24 heures[1].

L'autre conséquence de la très grande taille du fichier, sur laquelle on reviendra, a été la lourdeur de la mise en œuvre : quand une estimation (et encore, une fois terminées les phases de calcul des probabilités d'inclusion) sur quatre variables immobilise le micro pendant près de deux heures, on hésite à tester toutes les variantes souhaitables dans la description du plan de sondage, et à le faire sur la variété souhaitable de variables d'intérêt : les variables descriptives du handicap présentes dans HID ont en effet à la fois des fréquences très variées (par exemple, autour de 200 000 aveugles mais plus de 22 000 000 de personnes ayant au moins une déficience), et des dispersions très inégales parmi les groupes de sévérité qui déterminent les probabilités d'inclusion HID et donc les poids (aucun aveugle dans le groupe 1, celui où les poids sont très élevés, mais la présence dans ce même groupe d'un nombre suffisant de personnes souffrant de handicaps d'origine psychique pour qu'elles représentent la moitié du total de cette population « psychiquement handicapée »)..

[1] Petit détour : un bug de ce type sur un bon logiciel tout fait du type Microsoft est imparable : on n'a plus que les yeux pour pleurer et l'espace-temps pour attendre. Heureusement, il s'agissait d'un logiciel « maison », dont l'auteur était encore en poste ; on aurait sans doute eu une issue également favorable, quoique sans doute moins rapide, avec une application du monde Linux. Autant le dire quand ça se passe…

1.4.2 Les apports de Poulpe à la description du plan de sondage

L'application fournit tout d'abord de nombreuses aides dans les deux premières phases (ARBGEN et CALPII - voir 1.2) de description du plan de sondage et de calcul des probabilités d'inclusion subséquentes. On en citera deux :

- un message signale tous les cas où un seul élément (une seul zone de délégué, un seul district, un seul individu…) a été tiré dans l'une des étapes de sondage décrites. Cela évite d'avoir à contrôler à l'avance le nombre de districts tirés dans chaque zone de délégué, ou le nombre de répondants VQS dans chaque district. Naturellement, ces signalements permettent de repérer les corrections ou modifications à effectuer - à la charge de l'utilisateur car il n'y a pas de procédure de regroupement automatique dans Poulpe ;

- de la même façon, un message signale les cas où les probabilités d'inclusion sont supérieures à 1, par exemple les cas où on a plus de réponses dans un district selon le fichier VQS que le recensement n'avait enregistré d'habitants (ce problème survient dans l'étape de modélisation de la non réponse VQS au travers d'un sondage aléatoire simple). Il s'agit alors le plus souvent d'erreurs d'appariement dans la construction des fichiers, ou de dénomination du district impliqué, ou de report des effectifs du recensement. C'est un guide de correction précieux.

  Dans quelques cas le surnombre de répondants est très faible (un ou deux individus) et correspond à des individus réintégrés dans un autre logement - et un autre district - par le RP, mais ayant répondu à VQS dans ce district. Cette anomalie est sans conséquence sur les calculs menés par Poulpe.

1.4.3 Les apports de Poulpe à l'appréciation de la qualité de l'enquête

Le logiciel fournit tout d'abord les estimations de variance des estimateurs sans biais de Horvitz-Thompson (HT), celles des écarts-types et des intervalles de confiance. C'est naturellement ce qui était attendu. Le plus intéressant est sans doute la variété des coefficients de variation. On se reportera au tableau 5, dans lequel figurent une dizaine de variables présentant des caractéristiques assez différentes : prévalences allant de 0,2 % (MOB1) à 38,7 % (DEFI1) ; phénomènes étroitement liés à des pratiques administratives dont on sait qu'elles étaient géographiquement très inégales au moment de l'enquête (ALLOC1, COTOR1, INVAL1) versus des pratiques probablement plus dépendantes de caractéristiques socio-familiales (AIDKI1) ; difficultés encore souvent vécues comme « honteuses » et donc diversement sous-déclarées (EXPR1) versus des usages sans enjeu de représentation sociale particulier (DADAPT1). On trouvera au tableau 6 les coefficients de variation initiaux et ceux obtenus après les aménagements détaillés dans le second chapitre.

Globalement, après les corrections, les coefficients de variation se situent entre 1,3 et 4,3 %, ce qui est assez satisfaisant, sauf pour deux variables : MOB1 (CV = 6,4 %) variable exprimant le confinement au lit ou au fauteuil, dont l'estimation est probablement plus imprécise en raison de sa faible fréquence[2] (0,2 %) et EXPR1 (CV = 8,1 %), variable exprimant l'illettrisme, dont l'estimation est imprécise en raison des modalités de réalisation de l'enquête. En effet, le classement dans les groupes VQS (bâti d'après les réponses fournies à la pré enquête) détermine la probabilité de tirage et donc le coefficient de pondération. Par ailleurs, dans le domaine de l'illettrisme la déclaration d'une difficulté au moment du Recensement, sur un questionnaire rempli en dépôt-retrait, est probablement difficile. Une proportion non négligeable de personnes n'ayant que ce type de problème ne l'auront sans doute pas déclaré à VQS ; elles seront par voie de conséquence classées dans le groupe 1 et affectées d'un poids élevé. Une partie d'entre elles rectifiera sa déclaration lors de la procédure d'enquête HID, qui

[2] Le coefficient de variation d'une proportion p varie comme l'inverse de la racine carrée de p.n, où n représente la taille de l'échantillon.

se fait en face à face, au cours d'un entretien plus long, plus précis, plus convivial aussi sans doute. On aura donc de ce fait des personnes avec de légères déficiences intellectuelles s'exprimant par un problème d'illettrisme affectées d'un coefficient de pondération élevé - d'où une dispersion accrue des poids et un gonflement de la variance et de l'imprécision des estimations.

Ce type de problème se rencontre également pour d'autres difficultés mal déclarées à VQS, comme les problèmes liés à la santé mentale. On a par exemple observé par ailleurs que la moitié de l'estimation de l'effectif des personnes suivies régulièrement pour un problème de santé mentale provenait de personnes classées dans le groupe VQS 1. A titre d'exemple de l'effet inverse, on a estimé le coefficient de variation pour les personnes ayant une prothèse et pour celles ayant une prothèse des membres inférieurs ; alors que les prévalences sont assez modestes (2,2 % et 1,17 %) les coefficients de variation sont également assez peu élevés (respectivement 3,5 et 4,5 %). Il faut dire que la déclaration d'une telle situation est aujourd'hui dans l'état de nos relations socioculturelles beaucoup plus aisée que celle d'un problème intellectuel ou mental.

A côté des estimations de variance effectuées selon la description complète du plan de sondage, Poulpe calcule à la demande deux autres types d'estimations : (1) les variances « avant calage », qui tiennent compte de la description du plan de sondage, mais sans utiliser la pondération issue du calage de l'enquête (par CALMAR) sur des marges populationnelles issues en général du recensement; (2) les variances dites par « sondage aléatoire simple », calculées comme si l'échantillon final avait été tiré de manière aléatoire simple

Tableau 1: Estimations initiales de coefficients de variation, variances, effets de calage et de sondage

Variable d'intérêt	Variance de l'estimateur de HT du total	coefficient de variation (en %)	effet de calage	effet de sondage
AIDKI1	23 268 269 755	3,0	0,18	0,59
ALLOC1	30 645 469 026	7,8	0,53	1,25
CONFIN1	3 863 599 767	10,7	0,41	0,65
COTOR1	19 990 923 769	6,7	0,23	0,82
DADAPT1	1 837 940 274	5,0	0,39	0,26
DEFI1	219 976 186 130	2,1	0,06	2,43
EXPR1	36 366 676 550	13,9	0,27	2,92
HANDI1	318 425 419 990	3,1	0,09	3,76
INVAL1	84 341 974 284	8,3	0,55	2,56
MOB1	227 261 954	12,6	1,00	0,23
MOB2	1 771 495 406	4,3	0,21	0,20
MOB3	1 242 856 682	4,8	0,14	0,15

Rappelons que le redressement par CALMAR effectué sur l'échantillon des répondants HID a eu pour rôle à la fois de limiter les effets de biais dus à la non-réponse ainsi que le défaut de couverture et de limiter les fluctuations d'échantillonnage HID.

L'effet de calage présenté dans le tableau 1 est le rapport entre la variance après calage et la variance avant calage. Il permet de mesurer, pour chaque variable d'intérêt, à quel point le redressement a amélioré la précision des estimateurs.

Pour toutes les variables étudiées, l'effet de calage est inférieur à 1, sauf en ce qui concerne la variable Mob1 où il est égal à 1. Cela signifie que le fait d'avoir redressé l'échantillon a augmenté la précision de l'ensemble des variables d'intérêt. La moyenne géométrique sur l'ensemble des douze variables de l'effet de calage vaut 1/3.

Les variables dont les estimations de variance enregistrent le plus fort effet de calage sont les variables Defi1 (1/16) et Handi1 (1/11). *Le redressement a donc permis d'estimer le nombre de métropolitains souffrant d'au moins une déficience avec une précision égale à celle que l'on aurait eu sans redressement sur un échantillon de taille 16 fois plus grande !!!*

Une manière d'apprécier la qualité du plan de sondage HID, du point de vue de son effet sur la précision des estimations, consiste à se reporter à la dernière colonne du tableau 1. Poulpe y calcule « l'effet de sondage », rapport entre la variance qui découle du plan de sondage effectivement modélisé et la variance que l'on obtiendrait si on avait utilisé un sondage aléatoire simple de même taille (il s'agit des variances après calage).

Ce coefficient est la résultante de deux effets essentiels, plutôt antagonistes: le tirage à plusieurs degrés sur lequel s'appuie VQS génère de l'effet de grappe et tend donc à faire passer l'effet de sondage au-dessus de 1. Mais la post-stratification HID, en s'appuyant sur des catégories a priori homogènes du point de vue des variables qui nous intéressent, joue comme une stratification et tend à faire passer le coefficient en dessous de 1. Il faut néanmoins moduler ce principe parce que la grande disparité des taux de sondage de seconde phase complique le paysage (une très faible taille d'échantillon dans une catégorie assez nombreuse et néanmoins un peu hétérogène peut avoir un effet dévastateur).

Sept des douze variables d'intérêt présentent un effet de sondage inférieur à 1 : Mob1,2,3, Dadapt1, Confin1, Aidki1 et Cotor1. *Il est intéressant de constater que ce sont précisément les variables qui définissent la dépendance des personnes.* Sur l'ensemble des variables étudiées, l'effet de sondage a une moyenne géométrique de 0.77, donc la précision est globalement meilleure que celle que l'on aurait obtenue par sondage aléatoire simple.

Les variables Mob1,2,3 et Dadapt1 ont un effet de sondage inférieure à 0.25, ce qui signifie que le plan de sondage a permis d'obtenir une précision quatre fois plus grande que le sondage aléatoire simple. *Si les individus étaient tirés au hasard dans la population métropolitaine, il aurait fallu un échantillon quatre fois plus grand pour obtenir la même précision : une autre manière de commenter ce résultat consiste à dire que, pour ces variables au moins, la post-stratification permise par VQS s'est avérée très discriminante, et donc tout à fait efficace*

2. Recherche et correction des erreurs

Avoir réussi à décrire le plan de sondage et à faire produire des estimateurs de variance à Poulpe n'était qu'une première étape. Naturellement, il convenait ensuite d'interroger ces estimations sur leur… fiabilité.

On allait se heurter à une limite fortuite du logiciel et découvrir trois erreurs - deux mauvais choix dans la description du plan de sondage et une anomalie d'appariement - en analysant les estimations de Poulpe.

Donc les premiers résultats obtenus après ce «débuggage » avaient ceci de très déstabilisant : les estimations de totaux, ou de prévalences (proportions d'individus concernés par telle ou telle situation) faites par le logiciel étaient systématiquement supérieures (de près des deux tiers en moyenne) à celles obtenues par les pondérations des concepteurs - qui étaient aussi les plus vraisemblables. Par exemple, on obtenait une population totale de 88 millions d'habitants pour la France métropolitaine, ce qui était évidemment inacceptable.

2.2 Remise en cause du remplacement des Secteurs d'Agents Recenseurs par des Districts

L'un des atouts dont on a pu se servir tient à ce que dans son étape de «calcul des probabilités d'inclusion » (dite CALPII), le logiciel insère dans le fichier de données les probabilités élémentaires relatives à chaque étape du plan de sondage (tel qu'il a été décrit). On dispose ainsi de données permettant de vérifier l'adéquation de ces probabilités (globalement ou par sous-ensemble) : dilatent-elles par exemple correctement la population d'une étape à l'autre du plan de sondage ?

Ceci nous a permis de repérer une erreur assez importante dans la description du plan de sondage de la première partie (VQS). Elle concerne l'étape au cours de laquelle on a tiré au sein des zones de délégués des secteurs d'agents recenseurs. Rappelons d'abord deux éléments particulièrement gênants : (1) compte tenu des conditions imposées par les modalités de collecte du RP (emboîtement successif de zones de délégués puis d'agents recenseurs et nécessité d'inclure dans VQS l'ensemble du secteur de chaque agent recenseur concerné), on a été fréquemment amené à ne tirer qu'un agent recenseur VQS par zone de délégué ; (2) en tout état de cause, et malgré les demandes formelles adressées à la division « Recensements de la population » de l'Insee, aucune trace des secteurs d'agents recenseurs (ni leur définition géographique, ni l'appartenance des individus recensés à l'un ou l'autre des secteurs) n'a été conservée dans aucun des fichiers du recensement.

Le premier point entraîne une incapacité théorique dans les zones de délégués concernées à estimer la variance liée à cette étape du tirage - il eût fallu tirer au moins deux secteurs ; le second entraîne la nécessité de reconstituer au mieux un zonage susceptible d'intervenir dans le tirage.

Le choix initial avait été de prendre en compte les districts, unités plus petites que les secteurs d'agent recenseur (taille trois fois plus faible et nombre trois fois plus élevé en moyenne), et de faire « comme si » on avait tiré (de façon aléatoire simple) des districts et non des secteurs d'agent recenseur. L'avantage étant que les districts existent de façon stable, et sont conservés dans tous les fichiers du RP et des enquêtes collatérales (VQS ou enquête sur l'Étude de l'Histoire Familiale).

Ce faisant, on a omis de prendre en compte la très forte inégalité de taille des divers districts (y compris à l'intérieur d'une même zone de délégué). En modélisant leur sélection par un sondage aléatoire simple (méthode utilisée pour le tirage des secteurs d'agents recenseurs - qui sont eux de tailles proches), on a fortement, voire très fortement, surestimé la variance de cette étape et donc la variance globale expliquée par le tirage de VQS.

Pour rectifier la définition de cette étape, on a donc décidé de créer des secteurs d'agents recenseurs factices, de la façon suivante :

- on a d'abord classé de façon aléatoire les individus VQS de chaque zone de délégué ;

- on a ensuite considéré qu'il y avait deux secteurs factices dans les zones de délégués où en réalité un seul secteur avait été tiré (condition sine qua non pour pouvoir calculer une estimation de variance), et autant de secteurs factices que de secteurs réels dans les autres zones ;

- la population des répondants VQS a ensuite été répartie entre ces secteurs en quantités égales ;

- on a enfin considéré que ce sont ces secteurs qui avaient été tirés.

Par rapport aux estimations de variance provenant d'un tirage direct de districts, deux modifications de sens contraire sont intervenues : d'une part le nombre de secteurs étant inférieur au nombre de districts, cela a contribué à accroître la variance estimée ; d'autre part, la taille des secteurs d'une même zone de délégué étant très voisine et celle des districts le plus souvent très inégale, cela a contribué à diminuer l'estimation de variance.

Au total, comme le montre le tableau 1 ci-après, les estimations de variance calculées par Poulpe ont fortement diminué avec cette modification de la description du plan de sondage. Simultanément la surestimation des totaux et des prévalences a également décru de façon très significative. Pour en rester au même exemple, on est passé d'une population globale estimée de 88 millions à 66,7 millions.

Beau progrès, mais encore manifestement insuffisant. D'autant que cette surestimation s'amplifiait pour les variables à faible fréquence. Ainsi, alors que l'écart entre le population totale estimée par l'inverse des probabilités d'inclusion et celle résultant de l'application des pondérations de l'enquête n'était plus que de 16,5 % (au lieu des 52 % initiaux !), les écarts entre les estimations concernant la population utilisant une prothèse (DPRO) ou celle utilisant plus spécifiquement une prothèse des membres inférieurs (DPMI) - variables dont la prévalence est assez faible : 2,5 et 1,2 % - demeuraient considérables : respectivement 48 % et 52,5 %.

2.3 Redéfinition de la post-stratification préalable au tirage HID

On a donc été amené à revenir sur la description fournie à Poulpe de la partie du plan de sondage concernant le tirage de HID dans VQS.

Dans la réalité, le tirage de l'échantillon HID dans les répondants VQS a été effectué le plus souvent au niveau de chaque zone d'enquête (autour de 350 zones), partagée elle-même en 10 strates HID selon l'âge et l'indicateur résumé de handicap probable tiré des réponses à VQS (voir.1.1.2)

Les spécifications du modèle Poulpe n'acceptent au maximum que 99 « post-strates » de tirage pour cette deuxième phase du plan de sondage, au lieu des quelques 3500 utilisées (350 x 10). On avait donc dans un premier temps décidé de ne retenir que les 10 strates HID, croisées par le zonage « Hérault/Non Hérault », destiné à prendre en compte la très forte sur représentation de l'enquête dans l'Hérault.

Or on avait ce faisant omis les gros écarts de taux de sondage existant entre les 8 zones (7 départements et 1 région) « à extension VQS » (où VQS est sur représentée , mais pas HID) et les zones « sans extension VQS », où les probabilités de tirage de la phase HID sont évidemment plus élevées, puisqu'on tire des populations a priori comparables dans des bases de sondage de taille nettement plus petite..
On a donc convenu de croiser les 10 strates HID par un zonage en trois catégories « Pas d'extension / Extension VQS sans extension HID / Hérault »-puisque dans l'Hérault on a à la fois extension VQS et extension HID.

Ceci a eu un effet certain, mais curieux : l'effectif global calculé à partir de l'inverse des probabilités de sélection des répondants HID est revenu autour de 55 millions au lieu des 66 précédents, mais les

totaux des variables descriptives du handicap (et leurs prévalences) sont demeurés très supérieurs aux estimations de référence réalisées avec la pondération initiale présente dans le fichier des données d'enquête (POIDSCOR). Et ce, encore une fois, d'autant plus que les totaux (et les prévalences) étaient faibles.

Simultanément, cette correction a eu un effet sensible sur les estimations de variance, qui ont diminué de 40 à 60 %, au moins pour les trois variables pour lesquelles les calculs ont été effectués.

2.4 Correction d'une erreur d'appariement

En comparant, post-strate par post-strate, d'une part les estimations de totaux obtenues avec la pondération recalculée par le logiciel (en multipliant les probabilités de sélection à chaque étape du plan de sondage) et d'autre part les estimations de ces même totaux obtenues à partir des poids redressés présents dans le fichier d'enquête, on s'est rendu compte que la « surestimation » des tailles de population handicapées provenant du logiciel était intégralement imputable à la post-strate « 1 ». Celle-ci regroupe les individus de moins de 70 ans qui avaient déclaré n'avoir aucun problème à VQS : ils ont été tirés avec une faible probabilité, et sont donc affectés de poids très élevés (fréquemment proches de 50 000).

On a alors recherché quels étaient les individus de la strate 1 qui avaient ensuite déclaré de forts handicaps à HID (contribuant donc à une estimation élevée des effectifs de handicaps en raison de leurs poids élevés) et on s'est aperçu d'une erreur dans les procédures d'appariement : lorsqu'un ménage comprenait deux individus interrogés par HID, l'un handicapé et l'autre pas, on avait échangé par erreur les valeurs de la variable repérant la post-strate entre les deux individus en question. Du coup, des individus handicapés, qui auraient du être classés en post-strate 7, 8, 9 ou 10 et avoir une probabilité de tirage élevée et un poids calculé par le logiciel plutôt faible, se sont retrouvés classés à tort dans la post-strate 1, ce qui leur a valu, conformément à la description du plan de sondage l'attribution par le logiciel d'une probabilité d'inclusion faible et d'un poids élevé, approchant, atteignant ou dépassant la valeur de 50 000. D'où une surestimation considérable par le logiciel des diverses modalités « positives » du handicap.

On a procédé à la correction de cette (sérieuse) erreur et heureusement les écarts entre les estimations de totaux obtenues respectivement par le produit des inverses des probabilités d'inclusion et par les pondérations originelles de l'enquête sont devenus très faibles (de l'ordre de 1 à 2 %).

Tableau 2 : Estimateurs de totaux - L'effet des corrections successives

Variable d'intérêt	Estimateur du total (poids réels)	Estimation initiale de Poulpe	Estimation Poulpe après correction 1	Estimation Poulpe après correction 2	Estimation Poulpe après correction 3
ALLOC1	2 237 528	5 199 136	3 274 289	3 379 854	2 197 701
DEFI1	22 226 953	37 135 469	23 581 033	23 581 033	22 138 782
HANDI1	17 976 079	31 621 488	19 481 560	19 869 451	18 094 765
INVAL1	3 483 114	7 486 118	4 805 683	4 958 638	3 449 006

Simultanément, les estimations de variances ont à nouveau diminué, au point que l'écart avec les estimations initiales a été parfois considérable, allant de 24 % pour DEFI1 (variable dont la prévalence est la plus élevée, proche de 40 %) jusqu'à 80 % pour HANDI1 (prévalence de 31 %) 85 % pour ALLOC1 (fréquence de 3,9 %) et même plus de 96 % pour INVAL1 (fréquence de 6,1 %).

Tableau 3 : Estimateurs de variances - L'effet des corrections successives

Variable d'intérêt	Variance initiale	Variance après correction 1	Variance après correction 2	Variance après correction 3
ALLOC1	25 201 925 301	11 290 185 590	5 031 810 123	3 767 717 984
DEFI1	116 347 447 109	120 991 984 089	n.d.	88 860 139 576
HANDI1	520 931 169 484	190 957 788 064	113 739 551 114	103 485 240 191
INVAL1	95 832 313 467	32 257 562 822	16 586 710 181	3 438 247 578

2.4 Bilan des corrections du plan de sondage complet

In fine, cette expérience nous a conforté dans l'idée qu'un critère très efficace pour évaluer la pertinence de la description du plan de sondage est l'étude des écarts existants entre les estimations de totaux recalculées par le logiciel (pondérations inverses du produit des probabilités d'inclusion calculées à chaque étape), et les estimations originelles résultant de l'application des pondérations issues du redressement.

On trouvera ci-après trois tableaux présentant pour une douzaine de variables d'intérêt les principaux résultats des estimations, avant et après les trois corrections présentées ci-dessus. On notera que celles-ci diminuent très sensiblement les écarts-types et coefficients de variation (entre 10 à 80 %).

Tableau 4 : Estimations initiales sur douze variables

Variable d'intérêt	**Estimateur du total (poids réels)**	**Variance de l'estimateur de HT du total**	**écart type de l'estimateur HT**	**Borne inférieure (IC à 95 %)**	**Borne supérieure (IC à 95 %)**	**coefficient de variation (en %)**
AIDKI1	5 017 660	23 268 269 755	152 539	4 718 683	5 316 637	3,0
ALLOC1	2 237 528	30 645 469 026	175 058	1 894 414	2 580 643	7,8
CONFIN1	581 966	3 863 599 767	62 158	460 136	703 795	10,7
COTOR1	2 105 969	19 990 923 769	141 389	1 828 846	2 383 092	6,7
DADAPT1	851 930	1 837 940 274	42 871	767 902	935 958	5,0
DEFI1	22 226 953	219 976 186 130	469 016	21 307 681	23 146 225	2,1
EXPR1	1 370 555	36 366 676 550	190 700	996 783	1 744 329	13,9
HANDI1	17 976 079	318 425 419 990	564 292	16 870 067	19 082 091	3,1
INVAL1	3 483 114	84 341 974 284	290 417	2 913 897	4 052 331	8,3
MOB1	119 183	227 261 954	15 075	89 635	148 730	12,6
MOB2	982 243	1 771 495 406	42 089	899 748	1 064 738	4,3
MOB3	741 583	1 242 856 682	35 254	672 485	810 681	4,8

Tableau 5 : Estimations corrigées sur douze variables

Variable d'intérêt	Estimateur du total (poids réels)	Variance de l'estimateur de HT du total	écart type de l'estimateur HT	Borne inférieure (IC à 95 %)	Borne supérieure (IC à 95 %)	coefficient de variation (en %)
AIDKI1	5 017 660	18 917 637 357	137 541	4 748 079	5 287 241	2,7
ALLOC1	2 237 528	3 767 717 984	61 384	2 117 220	2 357 836	2,7
CONFIN1	581 966	500 798 272	22 379	538 104	625 828	3,8
COTOR1	2 105 969	1 645 679 501	40 567	2 026 458	2 185 480	1,9
DADAPT1	851 930	1 159 042 211	34 045	785 202	918 658	4,0
DEFI1	22 226 953	88 860 139 577	298 094	21 642 688	22 811 218	1,3
EXPR1	1 370 556	12 386 709 270	111 296	1 152 417	1 588 695	8,1
HANDI1	17 976 079	103 485 240 191	321 691	17 345 564	18 606 594	1,8
INVAL1	3 483 114	3 438 247 578	58 637	3 368 186	3 598 042	1,7
MOB1	119 183	57 295 806	7 569	104 347	134 019	6,4
MOB2	982 243	1 162 564 591	34 096	915 414	1 049 072	3,5
MOB3	741 583	1 037 511 629	32 210	678 451	804 715	4,3

Tableau 6 : Évolution de la précision (ou de écart type)

Variable d'intérêt	écart type initial	coefficient de variation initial	écart type après corrections	coefficient de variation après corrections	Évolution en %
AIDKI1	152 539	3,0	137 541	2,7	- 9,8
ALLOC1	175 058	7,8	61 384	2,7	- 64,9
CONFIN1	62 158	10,7	22 379	3,8	- 64,0
COTOR1	141 389	6,7	40 567	1,9	- 71,3
DADAPT1	42 871	5,0	34 045	4,0	- 20,6
DEFI1	469 016	2,1	298 094	1,3	- 36,4
EXPR1	190 700	13,9	111 296	8,1	- 41,6
HANDI1	564 292	3,1	321 691	1,8	- 63,0
INVAL1	290 417	8,3	58 637	1,7	- 79,8
MOB1	15 075	12,6	7 569	6,4	- 49,8
MOB2	42 089	4,3	34 096	3,5	- 19,0
MOB3	35 254	4,8	32 210	4,3	- 8,6

2.6 Résultats sur des prévalences dans des sous-populations

Un des usages privilégiés des estimations de variance, en particulier pour les études et recherches est naturellement de valider les écarts de mesure entre sous-populations. Par exemple, l'enquête HID estime que les hommes ont moins fréquemment des déficiences, et ont moins souvent besoin de l'aide régulière d'une autre personne dans les activités quotidiennes pour des raisons liées à la santé ; inversement, elle estime que les femmes bénéficient beaucoup moins souvent d'une allocation liée à un handicap.

Ces écarts a priori un peu contradictoires sont-ils statistiquement significatifs ? Poulpe permet de répondre par l'affirmative, comme le montrent les résultats présentés dans le tableau 7 : les intervalles de confiance à 95 % des estimations pour les hommes et pour les femmes sont disjoints sur chacune des quatre variables (suffixe H pour les hommes et F pour les femmes). L'opposition entre une plus grande fréquence des déficiences et du besoin d'aide chez les femmes et a contrario une reconnaissance sociale plus fréquente du handicap masculin est un phénomène réel, qui réclame des explications.

La même conclusion intervient pour des variables aux prévalences moyennes sensiblement plus faibles : l'utilisation d'une prothèse, ou plus précisément d'une prothèse des membres inférieurs est là encore plus fréquente chez les femmes, et l'écart est net et significativement différent de zéro à 95 %.

Tableau 7 : Comparaison de diverses prévalences selon le genre

Variable d'intérêt	Estimateur du total (poids réels)	écart type de l'estimateur HT	Borne inférieure (IC à 95 %)	Borne supérieure (IC à 95 %)	coefficient de variation (en %)
AIDKIH	6,61	0,30	6,03	7,19	4,5
AIDKIF	10,74	0,26	10,23	11,25	2,4
ALLOCH	5,34	0,21	4,92	5,75	4,0
ALLOCF	2,54	0,06	2,42	2,66	2,4
DPROH	1,89	0,117	1,66	2,12	6,2
DPROF	2,49	0,091	2,32	2,67	3,7
DPMIH	0,98	0,062	0,86	1,11	6,3
DPMIF	1,35	0,069	1,23	1,47	4,4

3. Un outil de base plutôt qu'une application finalisée

L'objectif du travail sur les estimations de variance de HID est de fournir aux équipes de recherche travaillant à l'exploitation de la base de données le moyen de calculer, de façon convenable et standardisée la précision des résultats qu'elles produisent.

Or l'enquête HID comporte plus de 500 variables d'intérêt brutes et les exploitations utiliseront en outre de nombreuses variables calculées - à commencer par les prévalences, tant sur l'ensemble que sur des sous-populations diverses, définies par de multiples critères socio-démographiques. Il n'est donc pas envisageable de produire toutes les estimations a priori, ni même de mener à l'Insee les calculs à la demande.

La seule solution est de fournir un outil suffisamment commode pour être utilisé par les demandeurs eux-mêmes. De ce point de vue, Poulpe constitue une base précieuse mais exige un travail d'habillage préalable.

3.1 Au minimum homogénéiser la description du plan de sondage

Des cinq étapes successives de l'application Poulpe (voir 1.2), les deux premières, qui sont consacrées à la description du plan de sondage et au calcul des probabilités d'inclusion, doivent impérativement être invariables d'une estimation et d'une équipe à l'autre. Leur mise au point nécessite, comme on l'a vu, une connaissance détaillée du plan de sondage que seule possède l'équipe conceptrice de l'enquête, et un travail de description comportant des approximations réfléchies sous le contrôle de méthodologues en sondages. Ce travail doit être effectué une fois pour toutes.

De par sa conception, Poulpe permet précisément de stocker les fichiers résultant de la seconde étape (CALPII), avec l'ensemble des identifiants et des probabilités d'inclusion. Il est donc à la fois souhaitable et envisageable de fournir sur cette base une application qui reparte de ces fichiers et ne comporte que les deux étapes finales définissant les variables pour lesquelles on demande les calculs et réalisant les estimations.

Le « verrouillage » des deux premières étapes présente un double intérêt :

- d'une part il assure que les estimations de variance ou d'intervalles de confiance effectuées par les diverses équipes sur une même variable seront identiques, ce qui est tout de même souhaitable ;

- d'autre part il économise à ces équipes le lourd travail initial de préparation informatique, qui comprend la conception et la réalisation des programmes de description du plan de sondage, la réponse aux nombreux messages du logiciel signalant des échantillons réduits à une unité, ou des probabilités d'inclusion supérieures à un, et enfin le déroulement de l'étape de calcul des probabilités d'inclusion (CALPII), dont la durée peut approcher une heure et demie de « temps CPU » - pour reprendre un vieux concept datant de l'époque « grosse informatique ».

3.1.1 Les éléments à fournir

Cinq éléments doivent être fournis aux équipes utilisatrices :

1. en premier, naturellement le logiciel compilé (il s'agit d'un ensemble de macros SAS dans lesquelles chaque « demande Poulpe » va puiser selon les besoins définis par l'utilisateur) ;

2. en second, plusieurs fichiers utilisés au cours des deux premières étapes de Poulpe - dont la présence est indispensable - doivent également être fournis, comme le fichier « arbre descriptif du plan de sondage », ou celui comportant les effectifs des unités concernées aux divers niveaux géographiques (« fichier géographique »), ou encore celui stockant le journal des exploitations réalisées…

3. en troisième, trois modèles d'appels de macros : celui correspondant à l'étape CHARLIS, qui définit les variables à partir desquelles vont être réalisées les estimations ; celui correspondant à l'étape ESTIVAR, qui produit les estimations de variance sur des variables simples ; celui correspondant à l'étape ESTIFON, qui produit par une technique de linéarisation les estimations de variance sur des « fonctions » (prévalences, sommes…) ;

4. en quatrième, figure le fichier individuel des données d'enquête, comprenant en particulier les identifiants et les probabilités d'inclusion nécessaires aux calculs et résultant de l'étape CALPII. Ce fichier ne peut prétendre comporter toutes les variables d'intérêt plausibles. La stratégie proposée consiste donc au contraire à le dépouiller au maximum, et à l'accompagner d'une macro ou d'un exemple de programme lui adjoignant à la demande les variables d'intérêt choisies par l'exploitant ;

5. enfin naturellement un fichier guide indiquant où disposer les divers éléments fournis et comportant le mode d'emploi des appels de macros, sous forme papier et informatique, doit accompagner la livraison.

On notera la souplesse de Poulpe pour préparer ce type de montage. Le logiciel fournit en effet, pour toute demande correspondant à l'une quelconque de ses étapes, un appel de macro, qui peut être recopiée sous forme de programme SAS, et lancée en tant que tel, avec ou sans modification.

D'autre part, il génère et stocke, à l'issue de chaque étape, un ou plusieurs fichiers fixant les résultats obtenus.

Aussi le montage d'une application spécifique à telle ou telle enquête sur la base du logiciel est-il particulièrement aisé.

3.1.2 Les limites de cette solution dans le cas d'HID

Pour séduisante qu'elle soit, cette solution comporte deux inconvénients liés au plan de sondage HID.

Le premier est un problème de lourdeur. Dans la mesure où HID est une enquête en deux phases, où l'échantillon de la seconde phase a été tiré parmi les répondants de la première phase VQS, les calculs de Poulpe se déroulent sur un fichier comprenant l'ensemble des observations de la première phase du plan de sondage. Soient 330 000 observations dans le cas présent. En effet, l'approche « standard » utilise des comptages dans le fichier de première phase pour calculer les taux de sondage utilisés dans les différentes post-strates.

Du coup, les temps de passage et les espaces de travail temporaires nécessaires sont considérables. Par exemple, on a tenté de mener un calcul simultané de variance sur une vingtaine de variables ; il a provoqué la constitution simultanée de trois fichiers temporaires de 1,5 giga-octets chacun ; naturellement, pour des raisons d'espace le programme a planté. Il n'est pas conseillé de mener un calcul simultané sur plus de 4 variables ; encore le temps de passation sur un PC PIII-800 avec 128 méga-octets de RAM approche-t-il les deux heures, pendant lesquelles le poste est quasiment inutilisable pour une autre tâche.

La seconde limite tient à ce que la deuxième phase de l'enquête ne peut, selon les spécifications de l'application, comporter plus de 99 post-strates. Et d'ailleurs, plus on augmente le nombre de strates et plus le volume des fichiers de travail augmente, ainsi que les temps de calcul. Pour prendre un exemple, si l'on voulait décrire qu'on a tiré l'échantillon HID dans chaque région, selon des taux dépendant des réponses à VQS en 10 catégories, soit un nombre de strates de l'ordre de 200 (20 régions * 10 catégories VQS)… on irait bien au-delà des limites autorisées.

3.2 La recherche d'un plan de sondage « tronqué »

Aussi était-on tenté de rechercher une description du plan de sondage permettant de ne mettre en œuvre dans les procédures d'estimations que la deuxième phase HID, et donc un fichier de taille beaucoup plus modeste (21 740 observations au lieu des 330 000 utilisées dans le plan complet).

3.2.1 Possibilité et utilité

Naturellement, ceci n'est acceptable que si le « court-circuitage » de la première phase VQS n'a qu'un effet modeste sur les variances estimées. Ce point avait été abondamment discuté à l'occasion du travail d'estimation sur petits domaines. L'avis intuitif, a priori, des méthodologues impliqués dans ce travail était que la part de variance imputable à la phase VQS devrait être sensiblement inférieure à celle imputable au tirage HID, le premier argument étant que, globalement, les effectifs des répondants VQS sont près de 20 fois supérieurs à ceux des répondants HID . Un second élément militant dans ce sens est que les allocations de seconde phase sont très éloignées d'allocations proportionnelles : les disparités de taux de sondage sont par principe et par choix extrêmement fortes dans la phase HID (dans un rapport de 1 à 100), les taux de sondage les plus faibles concernant un groupe représentant les trois quarts de la population globale. On doit se souvenir parallèlement que le tirage VQS génère certes des poids inégaux, mais néanmoins peu dispersés : de ce point de vue, le plan de sondage VQS est plutôt rassurant et cela renforce l'idée que la variance associée au tirage VQS devrait être très modeste.

Donc il apparaissait a priori envisageable d'utiliser pour les estimations de variance un plan de sondage approché négligeant la phase VQS, au prix évidemment d'une adaptation du système de pondération utilisé par le logiciel. On se retrouverait alors avec un plan de sondage fictif en deux phases (au lieu de trois) comportant le tirage de l'échantillon HID comme première phase et le processus de non-réponse comme deuxième phase.

Les avantages sont importants :

1. on travaille alors sur un fichier de données beaucoup plus petit ; la conséquence est immédiatement perceptible, car le temps de calcul de l' étape ESTIVAR tombe de près de deux heures à quelques minutes (moins de cinq !) ;

2. l'impossibilité de décrire le tirage HID en plus de 99 strates est levée, car cette limite technique du logiciel n'intervient que pour une deuxième phase. Notons que dans ces nouvelles conditions de temps de passation minimes, la souplesse de Poulpe peut alors donner toute sa mesure. On peut examiner aisément comment varient les estimations de variance selon le degré de simplification de la description du tirage : à partir de premières expériences, il semble qu'à partir du moment où on distingue les zones à extension VQS et l'Hérault, le raffinement consistant à augmenter le nombre de strates géographiques croisant les dix niveaux de réponse à VQS (pour se rapprocher de la réalité) joue peu. Que l'on distingue 30 strates (trois fois dix), 90 (neuf fois dix) ou même 340 (trente-quatre fois dix), les estimations de variance seraient pratiquement identiques.

3.2.2 Réalisation

Cette étape du travail est en cours. Elle demande en premier lieu une validation de la méthode de « court-circuitage » de VQS, laquelle ne va pas de soi. Sur le principe, il faut en effet modifier certaines informations présentes dans le fichier HID afin que les estimateurs de variance implantés dans Poulpe conduisent bien à des valeurs numériques proches de celles que fournit le logiciel avec la description complète du plan de sondage A titre indicatif, les premiers essais ont abouti à des

estimations de variances supérieures à celles obtenues pour le plan de sondage complet VQS-HID. Cela peut traduire deux situations : ou bien le plan de sondage fictif a été mal paramétré (il faut regarder de plus près cet aspect très technique),ou alors il faut envisager que l'intuition de base ne soit pas bonne, à savoir que la variance HID soit finalement relativement faible, en tout cas pas largement dominante dans le calcul de la variance d'ensemble ! Ce cas de figure dépend largement de la qualité des post-strates HID : s'il s'avère que VQS est une enquête filtre très efficace, alors les post-strates construites seront très homogènes et la variance associée à l'échantillonnage HID s'avérera faible.
Cela dit, si l'intuition s'avère exacte, le travail consistera à mettre en forme le produit simplifié à livrer aux équipes utilisatrices. Rappelons brièvement sa constitution :

1. le logiciel Poulpe compilé ;

2. trois modèles d'appels de macros : correspondant aux étapes CHARLIS, ESTIVAR et ESTIFON ;

3. le fichier individuel (21 740 observations) résultant de l'étape CALPII ;

4. les fichiers de travail du logiciel issus des deux premières étapes de Poulpe ;

5. et le mode d'emploi.

Conclusion :

De ces investigations, il ressort quelques points importants à garder en mémoire :

- Il ne faut pas perdre l'information qui a été utilisée au cours du processus d'échantillonnage, faute de quoi on n'est plus capable de calculer « proprement » une variance ;

- Il convient de décrire à Poulpe un plan de sondage aussi proche que possible de la réalité : des écarts entre réalité et modélisation - même d'apparence anodine - peuvent avoir des conséquences numériques fortes ;

- Les tirages dans lesquels la taille d'échantillon vaut un sont une source de problème pour le calcul de variance : on peut y penser au stade de la conception du plan de sondage.

- La manipulation de gros échantillons par Poulpe crée des difficultés multiples au niveau informatique

- L'examen des estimations de totaux obtenues avec les pondérations calculées par Poulpe est un élément de contrôle de pertinence très appréciable et très efficace.

Bibliographie

[1] « Logiciel Poulpe : guide de l'utilisateur », *Documentation technique de l'Unité Méthodes Statistiques,* INSEE, octobre 1998.

[2] Joinville O., « Mise en oeuvre du logiciel Poulpe pour estimer la précision de l'enquête « Handicaps-Incapacités-Dépendance » », *Rapport de stage,* Division Enquêtes et Études Démographiques, INSEE, septembre 2002.

[3] Mormiche P., «L'enquête HID de l'INSEE. Objectifs et schéma organisationnel», *Courrier des Statistiques*, 87-88, décembre 1998

[4] Des Raj., « Sampling theory », *Mac-Graw Hill*, 1968 (page 216)

Lexique des variables HID utilisées dans la note :

Aidki1 :	Etre aidé régulièrement par une ou des personnes dans l'accomplissement de certaines tâches de la vie quotidienne en raison d'un handicap ou d'un problème de santé
Aidkih (Aidkif) :	**Aidki1** pour les hommes (resp. pour les femmes)
Alloc1 :	Bénéficier d'une allocation en raison d'un problème de santé ou d'un handicap
Alloch (Allocf) :	**Alloc1** pour les hommes (resp. pour les femmes)
Confin1 :	Etre confiné dans son logement en raison d'un handicap ou problème de santé
Cotor1 :	Avoir déposé un dossier auprès de la Cotorep
Dadapt1 :	Disposer d'équipements spécialement adaptés à son handicap
Defi1 :	Avoir déclaré au moins une déficience au cours de l'interview
Dpmi :	Utiliser une prothèse des membres inférieurs
Dpmih (Dpmif) :	**Dpmi** pour les hommes (resp. pour les femmes)
Dpro :	Utiliser un appareil de remplacement d'une partie du corps (prothèse)
Dproh (Dprof) :	**Dpro** pour les hommes (resp. pour les femmes)
Expr1 :	Avoir plus de six ans et ne savoir pas lire, pas écrire ou pas compter
Handi1 :	Avoir déclaré à la première question de l'interview souffrir de gênes ou handicaps liés à un problème de santé
Inval1 :	Avoir une reconnaissance officielle ou assurantielle d'un taux d'incapacité ou d'invalidité
Mob1 :	Avoir un indicateur Bcolvez=1
Mob2 :	Avoir un indicateur Bcolvez=2
Mob3 :	Avoir un indicateur Bcolvez=3
Poidscor :	Coefficient de pondération issu du redressement de l'enquête effectué par les concepteurs
TaidF :	Proportion de femmes aidées dans certaines tâches de la vie quotidienne pour raison de santé ou handicap
TaidH :	Proportion d'hommes aidés dans certaines tâches de la vie quotidienne pour raison de santé ou handicap
TallF :	Proportion de femmes touchant une allocation en raison d'un problème de santé ou de handicap
TallH :	Proportion d'hommes touchant une allocation en raison d'un problème de santé ou de handicap
TinvF :	Proportion de femmes ayant une reconnaissance officielle ou assurantielle d'un taux d'incapacité ou d'invalidité
TinvH :	Proportion d'hommes ayant une reconnaissance officielle ou assurantielle d'un taux d'incapacité ou d'invalidité