UTILISATION DU LOGICIEL POULPE POUR LE CALCUL DE LA PRÉCISION D'ESTIMATEURS TIRÉS DE L'ENQUÊTE LOGEMENT 1996

David le Blanc

Introduction

Ce court article n'a pas de prétention méthodologique ; son ambition se borne à présenter l'application du logiciel POULPE à une enquête auprès des ménages.

Dans une première partie, on rappelle les particularités de l'enquête Logement par rapport aux autres enquêtes Ménages de l'Insee. La principale tient aux quantités que l'on cherche à estimer à partir de l'enquête : pour résumer, l'enquête Logement doit servir à estimer non seulement des structures, mais aussi des niveaux.

Le premier but de l'enquête Logement est de fournir un nombre de résidences principales, ou de ménages, mais aussi un nombre total de logements, les plus précis possibles. L'enquête sert également à donner des estimations du niveau de certaines variables portant sur le champ des ménages (ou des résidences principales) : nombre de propriétaires, de locataires privés et HLM, etc. Là encore, les niveaux ont de l'importance, car ces quantités doivent pouvoir être reliées à des grandeurs physiques, comme les flux d'aides à la pierre ou à la personne. Enfin, on calcule des estimateurs de type ratio, comme la part des propriétaires dans les ménages, de façon analogue à ce qui se pratique dans les autres enquêtes auprès des ménages.

La mise en oeuvre de POULPE suppose d'abord que l'on modélise le plan de sondage de l'enquête, afin de reconstituer les probabilités d'inclusion a priori de chacun des logements tirés. On doit ensuite modéliser la procédure de redressement et de calage. Vu la complexité du redressement de l'enquête, des simplifications sont nécessaires pour faire fonctionner POULPE.

Dans une deuxième partie, on s'intéresse aux résultats obtenus, selon deux optiques différentes :

- une perspective d'utilisation par les concepteurs de l'enquête. Quelle est la précision obtenue sur différents types d'indicateurs ? Comment cette précision

permet-elle d'interpréter les séries des enquêtes Logement depuis 12 ans ? Cette précision est-elle améliorée par le calage, et sur quelles variables ? Y a-t-il un « bon poids » ?

- une perspective plus méthodologique d'expertise de l'échantillon-maître et de la précision des différents types de variables dans les enquêtes tirées de cet échantillon. Il s'agit de calculer les précisions d'un grand nombre de variables, afin de pouvoir constituer des catégories, selon d'une part la perte de précision due au tirage dans l'EM (« design effect ») et d'autre part le gain de précision apporté par le calage sur marges (effet « CALMAR »). La comparaison de la précision des estimateurs calculée par POULPE avec celle qui résulte de calculs approchés permet en outre de donner des règles approximatives pour calculer la précision de variables quelconques, sans que le passage de POULPE soit nécessaire.

I) Plan de sondage et redressement de l'enquête Logement 1996

I-1) Plan de sondage

L'enquête Logement ne se distingue pas des enquêtes ménages traditionnelles au point de vue du plan de sondage. Deux bases de sondage sont utilisées : l'échantillon-maître (EM) et la base de sondage des logements neufs (BSLN). Le tirage s'effectue en une seule phase. Les taux de sondage sont les suivants :

Echantillon	Taux de sondage
EM :	
résidences principales et logements vacants non ruraux	1/ 720
logements secondaires ou occasionnels	1/ 1440
logements vacants en zone rurale	1/ 1080
BSLN (tous statuts):	1/ 361

I-2) Redressement

En revanche, en ce qui concerne le redressement, l'enquête Logement se distingue fortement des autres enquêtes ménages. Celles-ci sont généralement redressées en une seule étape, par calage de la structure de certaines variables de l'enquête sur les

marges de l'enquête Emploi la plus proche dans le temps. Ce calage est effectué sur des variables socio-démographiques comme le nombre de personnes, le nombre d'actifs, la strate géographique, la catégorie socioprofessionnelle et l'âge de la personne de référence du ménage. On considère en effet que l'enquête Emploi, avec un échantillon très important (plus de 100 000 logements), est la plus précise disponible à l'Insee ayant lieu à des intervalles de temps rapprochés.

Pour l'enquête Logement, on ne recourt pas à cette méthode, pour plusieurs raisons :

- un des résultats les plus attendus de l'enquête est l'évaluation du parc de logements et de ses différentes composantes (voir la **figure 1**). Il s'agit d'estimer un *nombre* de résidences principales, ou de ménages, mais aussi un nombre total de logements, les plus précis possibles. Ces quantités sont notamment utilisées dans une optique de suivi annuel du parc de logements et de ses différentes composantes, dans le compte satellite du Logement. Le fait qu'on cherche à relier les stocks de logements dans chaque catégorie aux flux annuels qui affectent ces catégories rend cruciale une bonne estimation du niveau des stocks, dans la mesure où les statistiques les plus fiables, celles du recensement de la population, sont très espacées dans le temps.

- le but premier des enquêtes Emploi n'est pas d'estimer précisément le parc de logements[1]. De fait, des divergences importantes existent entre niveau du parc de logements décrit par les enquêtes Emploi et celui donné par les enquêtes Logement. De plus, l'importance de l'échantillon de l'enquête Logement (40 000 logements) permet des raffinements en matière de redressement, de sorte que l'on peut penser que cette enquête estime le parc de logements aussi bien, sinon mieux, que l'enquête Emploi.

Des procédures de redressement spécifiques sont donc mises en oeuvre.

A) Correction de la non-réponse

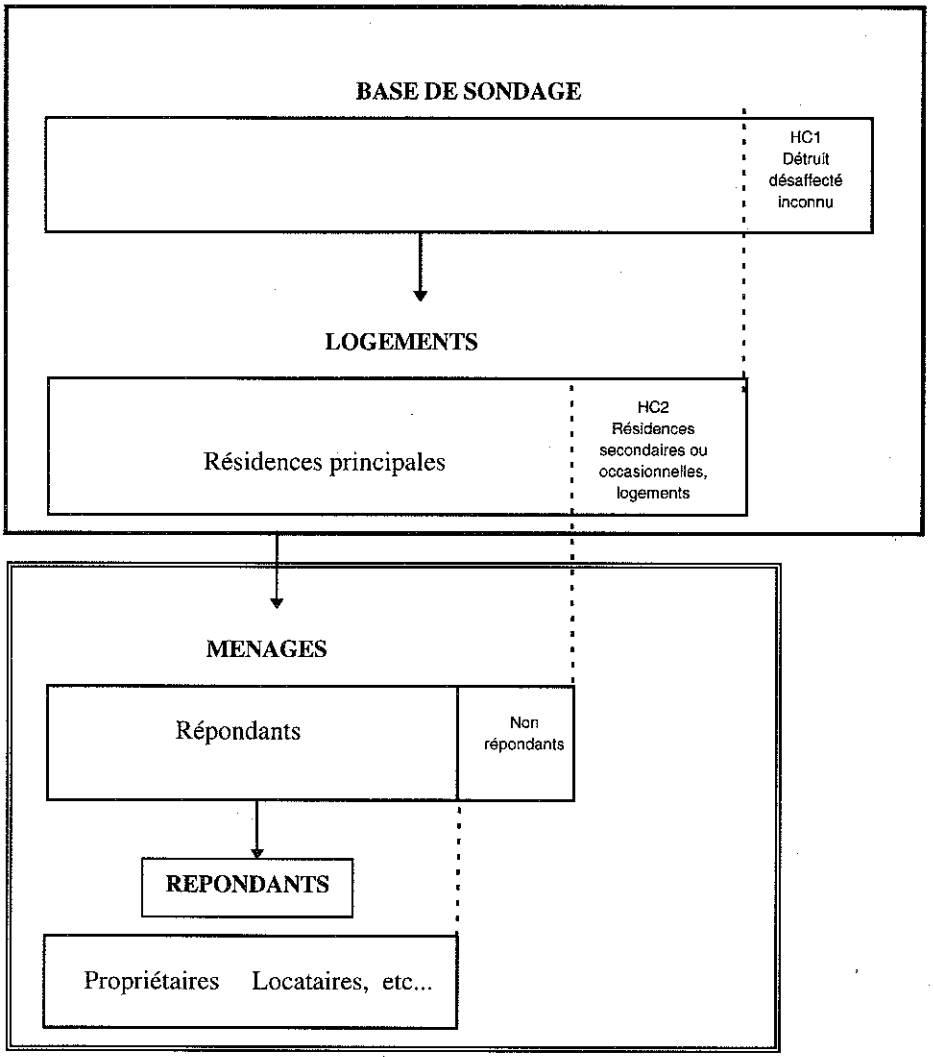
La non-réponse est corrigée à l'aide du logiciel CALMAR. En se plaçant sur le champ des résidences principales, il s'agit d'abord de déterminer les variables (disponibles dans la base de sondage) les plus discriminantes du point de vue de la non-réponse. Les poids des observations répondantes sont ensuite modifiés de manière à respecter les marges de ces variables évaluées sur l'ensemble des résidences principales.

1 L'enquête Emploi elle-même n'est pas calée sur un nombre exogène de ménages, mais sur la population estimée, stratifiée par sexe et âge.

Figure 1
Les différences entre l'enquête logement et une enquête ménage traditionnelle

Encadré noir : étape d'évaluation du parc de logements (spécificité de l'enquête Logement)

Encadré double : étape d'évaluation de variables ménages (toutes les enquêtes ménages)



Compte tenu des informations différentes disponibles dans les bases de sondage, on distingue trois sous-échantillons :

- les logements principaux,
- les logements non principaux au recensement de 1990,
- les logements neufs.

Sur chaque sous-échantillon, les variables les plus discriminantes pour la non-réponse sont mises en évidence à l'aide d'un modèle LOGIT (voir la **figure 2**). La macro CALMAR sert ensuite à modifier les poids des observations répondantes.

B) Calage sur les marges : correction des aléas d'échantillonnage

La deuxième étape consiste à caler l'échantillon (complet cette fois, c'est-à-dire l'ensemble des logements quelle que soit leur utilisation) sur des marges externes afin de limiter au maximum les fluctuations d'échantillonnage. Les logements tirés de l'échantillon-maître sont calés à partir de variables enregistrées au recensement de la population, les logements issus de la BSLN sur des marges tirées du fichier SICLONE des permis de construire.

Les variables qui servent au calage ne sont pas les mêmes selon que le logement est tiré d'une des deux bases de sondages, et pour l'échantillon-maître selon que son statut au recensement était résidence principale ou non. Ces variables sont indiquées dans la **figure 2**.

C) Correction des effectifs des bases de sondage

Cette étape est destinée à corriger la non-exhaustivité des bases de sondage. D'une part, il faut tenir compte des réaffectations de locaux en logements survenues depuis le dernier recensement. D'autre part, afin de tenir compte des données les plus récentes sur la construction neuve (enregistrées à partir des permis de construire), les poids des logements neufs sont modifiés.

La procédure de redressement suivie peut sembler inutilement lourde ; les résultats donnés par POULPE permettent de quantifier son efficacité, c'est-à-dire le gain de précision qu'elle apporte. Comme on va le voir, la précision obtenue sur des variables-clés (nombre de logements et de ménages, nombre de propriétaires) justifie tout-à-fait cette procédure de redressement par rapport à l'alternative qui consisterait à se caler sur des marges de l'enquête Emploi.

I-3) Application du logiciel POULPE à l'enquête Logement

La complexité du redressement effectué nécessite une simplification pour la mise en oeuvre de POULPE : le plan de sondage est approximé en particulier pour la base de sondage des logements neufs, ainsi que les procédures de correction de la non-réponse et de calage sur marges (voir la **figure 2**).

A) Deux niveaux de calculs de précision

Les particularités présentées au I-2) font que deux types d'estimation de la précision sont nécessaires, correspondant aux deux niveaux de variables produites par l'enquête :

1) précision sur le nombre de logements des différentes catégories (résidences principales, secondaires ou occasionnelles, logements vacants). La base de sondage étant supposée exhaustive, les hors-champ sont uniquement constitués par les logements inconnus ou ayant disparu. Ce cas ne sera pas détaillé ;

2) précision sur des variables de type ménage. Ces variables sont estimées sur le champ des résidences principales uniquement. On est donc dans un cas d'application de POULPE similaire à ce que l'on peut rencontrer dans les autres enquêtes ménages. Rappelons que l'idée sous-jacente du calcul de la précision de ces variables est de se ramener à des variables artificielles dont l'estimateur de Horvitz et Thompson corresponde à l'estimateur utilisé après redressement.

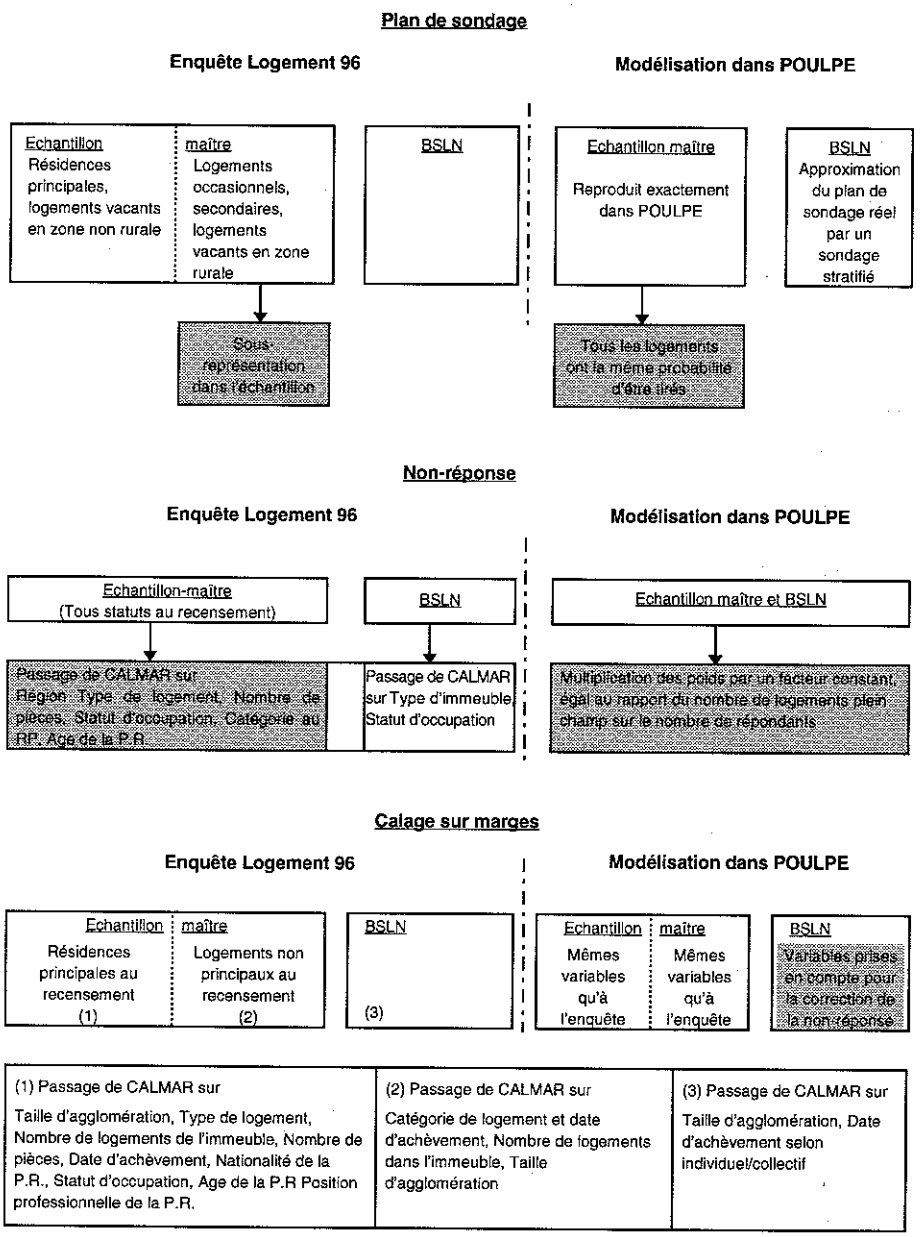
B) Précision sur des variables de type ménage

En ce qui concerne le **plan de sondage**, on considère que les logements issus de l'échantillon-maître ont tous la même probabilité d'être tirés, ce qui revient à négliger la sous-représentation des logements non-principaux. Pour les logements neufs, le plan complexe de la BSLN est approximé par un sondage stratifié, avec 33 strates (pour plus de précision, on se reportera au document méthodologique à paraître).

En ce qui concerne **le redressement**, deux approximations sont faites.

1) On considère que la non-réponse a été corrigée de façon globale, en multipliant les poids des observations répondantes par un facteur constant égal au rapport du nombre de résidences dans le champ de l'enquête sur le nombre de répondants. Pour pouvoir faire des calculs tenant compte de la correction de la non-réponse, on considère que cette correction constitue une phase supplémentaire dans le tirage de l'échantillon : les probabilités d'inclusion seront calculées comme si à partir de l'échantillon de logements, on avait tiré des résidences principales. Le logiciel POULPE sera donc paramétré comme pour un tirage en deux phases.

Figure 2
Plan de sondage et redressement de l'enquête Logement
Modélisation dans POULPE pour l'estimation de précision de variables ménages



(1) Passage de CALMAR sur	(2) Passage de CALMAR sur	(3) Passage de CALMAR sur
Taille d'agglomération, Type de logement, Nombre de logements de l'immeuble, Nombre de pièces, Date d'achèvement, Nationalité de la P.R., Statut d'occupation, Age de la P.R Position professionnelle de la P.R.	Catégorie de logement et date d'achèvement, Nombre de logements dans l'immeuble, Taille d'agglomération	Taille d'agglomération, Date d'achèvement selon individuel/collectif

A partir de ces données d'une part, et d'une modélisation de l'échantillon-maître et de la BSLN d'autre part, POULPE reconstitue le plan de sondage de l'enquête, en calculant les probabilités d'inclusion de tous les logements. Les probabilités a priori d'inclusion simple sont estimées en modélisant le plan de sondage de l'échantillon-maître par un arbre[2]. (se reporter à l'article de J.N. Petit).

Cette approximation de la non-réponse est évidemment grossière, et conduit sans doute à surestimer le gain de précision dû à la correction des biais d'échantillonnage.

2) On considère que le calage sur marge a été effectué indépendamment sur les trois sous-populations suivantes :

- logements principaux au recensement de 1990,
- logements non principaux au recensement de 1990,
- logements issus de la BSLN.

Pour les deux premières sous-populations, les variables utilisées dans POULPE sont celles qui interviennent dans le calage effectué pour corriger les aléas d'échantillonnage. Pour les logements neufs, les variables utilisées sont celles qui sont utilisées pour la correction de la non-réponse.

C) Résultats obtenus en sortie de POULPE

Pour chaque variable d'intérêt, trois calculs de précision sont réalisés :

① la précision obtenue sur les données brutes corrigées par un facteur constant correspondant au rapport du nombre de résidences dans le champ de l'enquête sur le nombre de répondants (simulation de correction de la non-réponse) et par conséquent avant le passage de CALMAR,

② la précision obtenue sur les données corrigées de la non-réponse et des fluctuations d'échantillonnage, c'est-à-dire après le passage de CALMAR ,

③ la précision obtenue sur les données corrigées de la non-réponse et des fluctuations d'échantillonnage en considérant que le plan de sondage est celui d'un sondage aléatoire simple (SAS). Ce calcul permet de comparer la variance obtenue par le plan de sondage complexe de l'échantillon-maître à celle que l'on aurait

---

2 Il faut noter que les probabilités d'inclusion simple sont calculées lors du tirage de l'échantillon des enquêtes ménages ; une amélioration évidente à apporter au calcul de précision des enquêtes consisterait à garder cette information pour alimenter le logiciel POULPE. On éviterait ainsi un calcul approximatif a posteriori. En pratique cependant, on peut vérifier que ces calculs donnent des résultats satisfaisants : les probabilités d'inclusion simple calculées par POULPE sont distribuées autour de leur valeur théorique.

obtenue si le plan de sondage avait été celui d'un SAS. Le rapport des deux variances estimées est appelé « design effect ».

Un moyen commode pour comparer le degré de précision de différentes variables consiste à comparer leurs coefficients de variation, sans unité. Pour évaluer le gain de précision dû à la correction de la non-réponse et au calage sur marges, on peut privilégier le rapport des variances avant et après le passage de CALMAR. Ce rapport peut en effet être interprété comme le gain réalisé en terme de taille d'échantillon, et donc de coût de l'enquête, pour un niveau de précision donné.

II) Résultats

Lors de l'examen des résultats, il importe de garder à l'esprit que les écarts-types estimés après calage sur marge calculés par POULPE sont basés sur l'hypothèse que les marges de calage sont connues sans erreur. Dans la pratique, cette hypothèse est couramment admise pour les chiffres tirés du recensement de la population ; pour ce qui est des logements neufs, cela est moins évident, dans la mesure où les statistiques tirées du fichier SICLONE du ministère de l'Equipement subissent des révisions importantes pendant plusieurs années après la date d'enquête.

II-1) Une optique « concepteur d'enquête »

A) Nombre de logements des différentes catégories

Les résultats numériques figurent dans le **tableau 1**. Ils confirment des idées déjà connues, mais que l'on ne pouvait quantifier auparavant.

- Le calage effectué lors du redressement a un effet important sur la précision de la mesure du parc de logements. L'écart-type sur le nombre de résidences principales est divisé par deux lors du redressement, ce qui signifie que la précision après redressement est celle d'un échantillon non redressé quatre fois plus grand.

- Le plan de sondage de l'échantillon-maître est très bon pour estimer les résidences principales, champ ordinaire des enquêtes ménages. Le design effect estimé après passage de CALMAR pour cette variable, d'une valeur de 1,23 , indique que la perte de précision par rapport à un sondage aléatoire simple est minime. Le nombre de ménages est connu (à 95 % de confiance) à plus ou moins 0,4 % près, ce qui représente environ 95 000 ménages. A titre de comparaison, la précision donnée par les concepteurs de l'enquête Emploi est de plus ou moins 105 000 ménages, pour un échantillon nettement plus important. La procédure de redressement sophistiquée est donc justifiée.

Tableau 1
Estimations de la précision par POULPE pour l'enquête Logement 1996. Nombre de logements selon le statut

	Estimateur pondéré (milliers)	Ecart-type avant calage	Ecart-type après calage	Coefficient de variation après calage (%)	Design effect après calage
Nombre de résidences principales	23 286	106 893	48 208	0,21	1,23
Nombre de résidences secondaires	2 452	104 218	41 935	1,71	2,68
Nombre de résidences occasionnelles	252	14 998	14 633	5,81	1,34
Nombre de logements vacants	2 231	41 444	39 075	1,75	1,23
Nombre de résidences secondaires et occasionnelles	2 704	105 173	42 652	1,58	2,31
Total des logements	**28 221**	**53 864**	**20 155**	**0,07**	**1,07**

- L'échantillon-maître se révèle nettement moins performant pour estimer le nombre de résidences secondaires. Le plan de sondage de l'échantillon-maître n'est pas optimal pour estimer le nombre de ces résidences, qui sont très concentrées géographiquement et conduisent à un fort effet de grappe (et donc à une variance importante). Toutefois, le passage de CALMAR permet de diviser par 2,5 l'écart-type sur cette estimation.

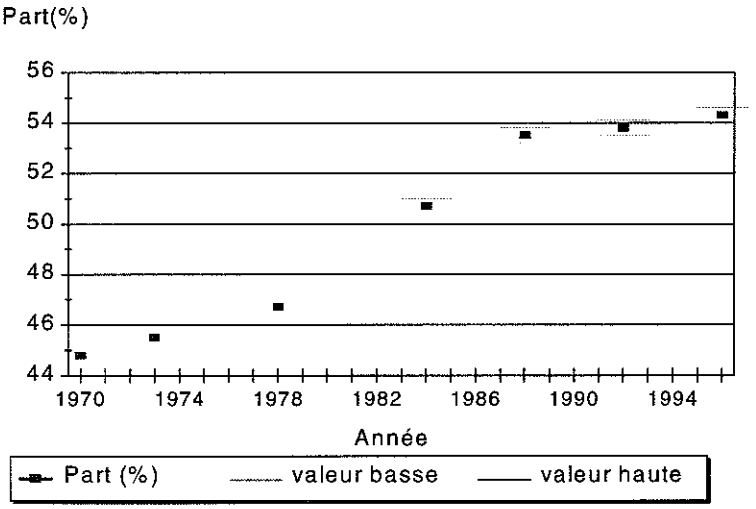
B) Variables ménages : parts des grands statuts d'occupation, flux quadriennaux

Indépendamment de la valeur intrinsèque d'estimations de précision pour une enquête donnée, disposer d'une estimation de la précision des données est nécessaire pour une exploitation correcte de la série des enquêtes Logement, particulièrement pour deux types de variables :

- la part des ménages dans les différents statuts d'occupation : propriété, location HLM, location libre. Un des résultats majeurs de l'enquête de 1992 avait été de mettre clairement en évidence que la part des ménages propriétaires, après une décennie d'augmentation rapide, demeurait stable, aux alentours de 54 %. Compte tenu de l'importance accordée à l'accession à la propriété dans les politiques du logement en France, ce résultat avait provoqué un certain émoi. Grâce à POULPE,

on sait que le passage de 53,8 à 54,3 entre 1992 et 1996 traduit plutôt une stabilisation qu'une reprise[3] (voir **graphique 1**) ;

Graphique 1
Evolution de la part des ménages propriétaires de leur résidence principale

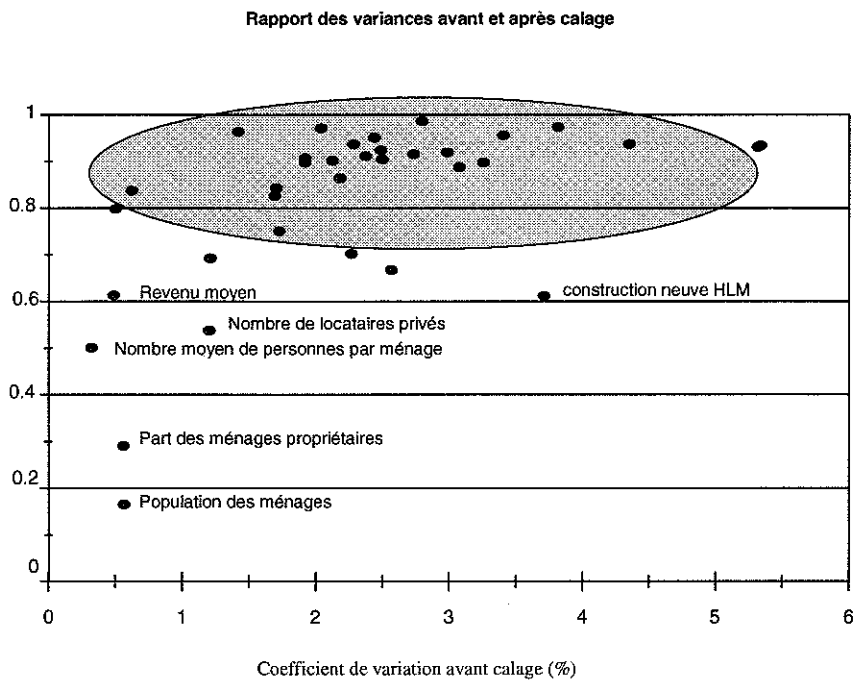


- les flux quadriennaux. Il s'agit de quantités mesurées à l'enquête 1996, portant sur des événements intervenues depuis la date de l'enquête précédente (novembre 1992). Ces flux permettent de mettre en cohérence les flux et les stocks entre les quatre dernières enquêtes. Dans la perspective d'utiliser la série des enquêtes Logement, qui ont lieu tous les quatre ans, pour suivre des générations d'une enquête à l'autre, on souhaite de même connaître la précision sur les tranches d'âge quadriennales.

3. La mise à disposition du logiciel POULPE a provoqué un regain d'intérêt pour les estimations de précision à l'enquête Logement. Celles-ci étaient avant 1992 tombées en désuétude, alors même que dans le passé des estimations avaient été produites (pour l'enquête de 1973 notamment). Il est frappant de rapprocher le besoin de précision de l'enquête et l'évolution de la part des ménages propriétaires, chiffre certainement le plus « sensible » de l'enquête. En 1973, la hausse de cet indicateur par rapport à 1970 était très faible (+ 0,7 point) : cette hausse traduisait-elle une augmentation réelle ? Par la suite, la progression rapide de la propriété (+ 0,54 point par an en moyenne) faisait apparaître entre deux enquêtes successives une différence évidemment significative, qui ne nécessitait pas de calculs de précision. Ce n'est qu'en 1992, au vu d'une stagnation sur les quatre dernières années, que la question s'est de nouveau posée avec acuité. Entre-temps, les « calculs approché » avaient sans doute perdu des adeptes.

Une des questions sous-jacentes est celle de la nécessité d'opérer des redressements particuliers pour réaliser de telles exploitations : en effet, les données de flux ne font pas partie des variables calées, car on ne dispose pas de marges externes pour ces quantités. En revanche, il serait tout-à-fait envisageable de caler les effectifs des ménages par tranche d'âge sur des estimations de population exogènes.

Graphique 2
Amélioration apportée par le calage sur marges



Les estimations de précision concernant ces variables sont données dans **le tableau 2**. Si l'on met à part le cas de la construction neuve HLM pour laquelle l'estimation donnée par POULPE n'est pas bonne (voir plus bas), les trois autres variables de flux ont des « design effects » proches de 1 ; le passage de CALMAR n'apporte qu'un faible gain de précision. Le cas des variables indicatrices des tranches d'âge est significatif : la tranche des ménages de plus de 65 ans, proche des modalités de la variable de calage (plus de 75 ans), voit son écart-type divisé par deux après le passage de CALMAR. En revanche, les tranches d'âge 20-24 ans et 24-28 ans étaient incluses à l'intérieur d'une même modalité de la variable de calage (moins de 30 ans) : leur précision n'est pratiquement pas améliorée par le passage de CALMAR. Finalement, si l'on veut travailler sur des tranches d'âge quadriennales, il faudrait sans doute caler sur des marges respectant ces tranches.

Tableau 2
Estimation par POULPE de la précision pour des variables de type flux et des tranches d'âge quadriennales.

Variable	Valeur (milliers)	Design effect CALMAR	Coefficient de variation CALMAR	ρ*
Nombre de nouveaux ménages	2 185	0.99	1.57	0.85
Construction neuve HLM	268	0.56	2.91	0.61
Nombre d'accédants récents	1 658	0.98	1.82	0.90
Nombre de ménages ayant changé de logement dans la même commune depuis 1992	2 570	1.07	1.54	0.83
Nombre de ménages de plus de 65 ans	5 642	1.06	0.64	0.32
Nombre de ménages entre 20 et 23 ans	640	1.00	3.09	0.90
Nombre de ménages entre 24 et 27 ans	1 264	1.09	2.17	0.90

* ρ : rapport des variances après calage et avant calage. C'est une mesure de l'amélioration apportée par le calage.

C) Conclusion

Le redressement compliqué ne sert « qu'à » améliorer la précision des variables vitales pour l'enquête (nombre de logements et de résidences principales, part des propriétaires, etc.). Il faut noter que ce constat justifie dans une large mesure la pratique courante de calage des enquêtes ménages (du moins celles dont le logement n'est pas l'objet principal) sur un nombre de ménages exogène. L'effet du redressement sur d'autres variables est beaucoup plus limité, en particulier sur les flux quadriennaux qui constituent une originalité de l'enquête Logement.

II-2) Une analyse plus globale

Le choix des variables pour lesquelles on a utilisé POULPE n'est pas innocent. Dans une perspective d'analyse des résultats, il s'agissait de représenter différents types de variables : totaux de variables indicatrices et ratios correspondants (par exemple, nombre de ménages locataires et part des ménages locataires), variables de type financier (moyennes de revenu ou de loyer), ratios portant sur un sous-champ (par exemple, part des locataires évoluant en secteur HLM), modalités très peu fréquentes dans la population étudiée.

Il s'agissait aussi de constituer des groupes de variables selon leur précision, avant et après calage, l'idée étant, pour chaque type de variables ainsi déterminé, de

rechercher des règles approchées permettant le calcul de la précision d'une variable quelconque, POULPE étant encore coûteux en temps et en espace disque pour le moment.

De cette analyse, on peut tirer deux enseignements principaux.

1) La variance estimée par POULPE pour des totaux de variables indicatrices et les ratios correspondants, qu'ils portent sur l'ensemble des ménages ou sur un sous-champ, peut être très correctement remplacée par une estimation approchée ne prenant pas en compte le plan de sondage.

On s'intéresse à une caractéristique des ménages. Soit X la variable indicatrice relative à cette caractéristique, et **X** son total dans la population des ménages. Soit p la proportion de ménages ayant la caractéristique étudiée. Le statisticien estime **X** et p par les estimateurs pondérés :

$$\hat{X} = \sum_r w_i X_i \quad , \qquad \hat{p} = \frac{\sum_r w_i X_i}{\sum_r w_i}$$

où r désigne l'échantillon des répondants et $(w_i, i \in r)$ est le jeu des poids des observations répondantes.

Supposons maintenant que l'on soit dans le cas d'un sondage aléatoire simple à probabilités égales dans un échantillon r de taille n tiré d'une population de taille connue N. On suppose que toutes les observations sont répondantes. Notons $Y = \sum_r X_i$ le total de la variable d'intérêt dans l'échantillon.

L'estimateur non pondéré de la proportion p est $\tilde{p} = \frac{Y}{n}$. Sa variance asymptotique $\tilde{p}$ est estimée simplement par , $\hat{V} = \frac{\tilde{p}(1-\tilde{p})}{n}$ et le coefficient de variation associé par $\tilde{C} = \frac{\sqrt{\tilde{V}}}{\tilde{p}} = \sqrt{\frac{(1-\tilde{p})}{n\tilde{p}}}$.

Le total **X** correspondant est estimé de la même manière par $\tilde{X} = N\tilde{p} = \frac{N}{n}Y$, des estimateurs de la variance et du coefficient de variation sont donnés par :

$$V' = \left(\frac{N}{n}\right)^2 Y(1-\frac{Y}{n}) = \left(\frac{N}{n}\right)\tilde{X}(1-\frac{\tilde{X}}{N}) \qquad \text{et}$$

$$C' = \sqrt{\frac{1-Y/n}{Y}} = \sqrt{\frac{N(1-\frac{\tilde{X}}{N})}{n\tilde{X}}}\,.$$

On propose donc les estimateurs approchés suivants pour le coefficient de variation :

a) Total d'une variable indicatrice X

$$C_t = \sqrt{\frac{N(1-\frac{\hat{X}}{N})}{n\hat{X}}} \quad \text{avec}$$

$\hat{X}$ total estimé de la variable X dans la population
n nombre de répondants à l'enquête
N nombre de ménages estimé à l'enquête

On peut encore écrire $C_t = \sqrt{\frac{1-\hat{p}}{n\hat{p}}}$, où $\hat{p}$ est l'estimateur pondéré de la proportion de ménages possédant la caractéristique étudiée.

b) Proportion de ménages possédant une caractéristique donnée

La caractéristique étudiée est toujours décrite par l'indicatrice X.

$$C_p = \sqrt{\frac{(1-\hat{p})}{n\hat{p}}} \quad \text{avec}$$

n nombre de ménages répondants à l'enquête
$\hat{p}$ estimateur pondéré de la proportion p dans la population

Cet estimateur est le même que celui du coefficient de variation du total de la variable X.

c) Proportion de ménages appartenant à un sous-champ de la population totale, possédant une caractéristique donnée

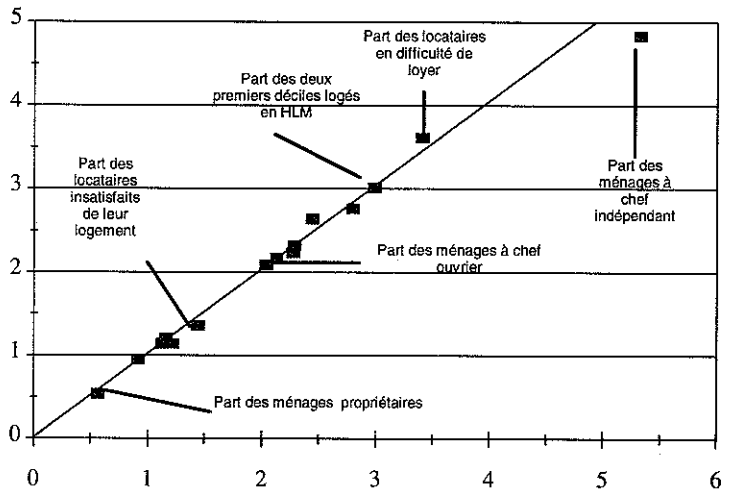
$$C_t' = \sqrt{\frac{(1-\hat{p})}{n_1\hat{p}}}$$

avec

n_1 nombre de ménages dans l'échantillon appartenant au sous-champ étudié
$\hat{p}$ estimateur pondéré de la proportion p dans la population

Les **graphiques 3 et 4** montrent que ces coefficients de variation approchés s'écartent très peu des valeurs calculées par POULPE pour les variables qui n'ont pas été calées. En revanche, pour les variables corrélées aux variables de calage, la précision donnée par POULPE est meilleure.

Graphique 3
Comparaison des coefficients de variation estimés par
avec des coefficients de variation approchés tirés des formules du II-2)
- Cas des proportions



Ainsi, pour exploiter l'enquête Logement, POULPE n'est pas nécessaire pour estimer l'ordre de grandeur de la précision d'une variable indicatrice quelconque : une formule approchée donne un résultat satisfaisant. Ce résultat n'est sans doute vrai que parce que l'échantillon de l'enquête est suffisamment important ; il n'est certainement plus valable pour d'autres enquêtes dont l'échantillon est plus réduit.

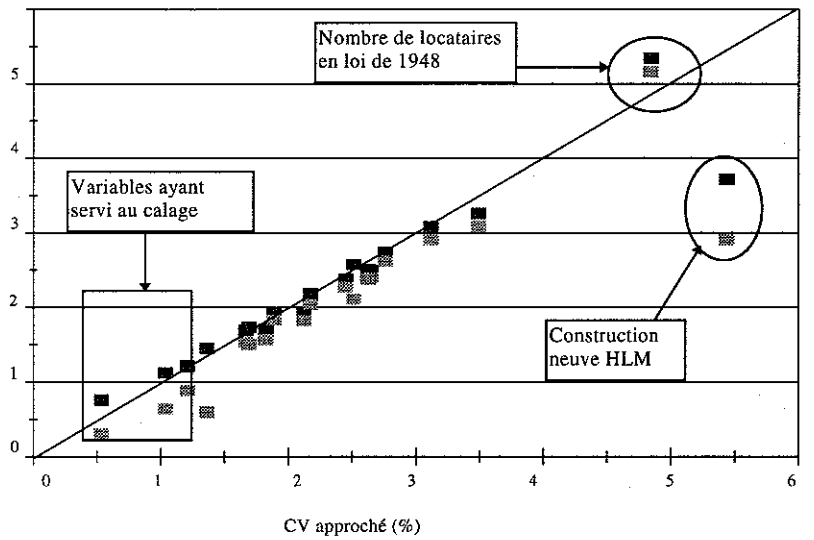
2) Le calage sur marge améliore de façon importante la précision des variables corrélées aux variables ayant servi au calage. Sur les autres variables, le gain de précision est négligeable.

Les graphiques 1 et 2 montrent que les coefficients de variation calculés par POULPE avant et après le passage de CALMAR diffèrent très peu, lorsque la variable étudiée n'a pas servi au calage.

En revanche, pour les variables ayant servi au calage, le gain de précision dû à CALMAR est important. Le coefficient de variation après calage est nettement inférieur à la valeur avant calage et aussi à la valeur approchée.

Graphique 4
Comparaison des coefficients de variation estimés par POULPE, avant et après calage, avec des coefficients de variation approchés tirés des formules du II-2) - Cas des effectifs

CV POULPE avant calage (noir) et après calage (gris) (%)



CV approché (%)

De façon plus anecdotique, on note que pour certaines variables importantes, l'estimation de la proportion des ménages concernés est meilleure (au sens du coefficient de variation) que l'estimation des effectifs. Cela signifie que les erreurs commises au numérateur et au dénominateur sont positivement corrélées. L'exemple le plus net est donné par la proportion de ménages propriétaires, avec un coefficient de variation après calage de 0,17, contre 0,31 pour le nombre de propriétaires. Toutefois, pour la grande majorité des variables étudiées, les coefficients de variation du total et de la proportion sont très proches.

I-3) Efficacité du plan de sondage et limites des outils actuels

Les résultats produits par POULPE permettent de juger de l'efficacité du plan de sondage de l'enquête, en même temps que de la perte ou du gain de précision dus à l'échantillon-maître. Pour la plupart des variables, le design effect estimé par

POULPE est très proche de 1 : cela prouve que l'échantillon de l'enquête a été correctement tiré.

Quelques cas surprenants méritent d'être mentionnés :

- les propriétaires et la population des ménages. Avant calage, le « design effect » est très fort (plus de 2), même si le calage a pour effet de le ramener au voisinage de 1. Or, ces variables sont censées être bien couvertes par l'échantillon-maître ;

- tout ce qui concerne le secteur HLM (hors construction neuve). Le calage améliore de façon importante la précision des estimateurs ;

- les loyers moyens en HLM et dans le secteur privé. Le design effect calculé par POULPE est très faible (entre 0,2 et 0,3).

Par ailleurs, la modélisation actuelle du plan de sondage de la BSLN (par un sondage stratifié) se traduit par une sous-estimation de la variance pour toutes les variables fortement corrélées avec la construction neuve. L'illustration de cette limite est visible sur le **graphique 4**. La variable CNHLM, qui représente l'effectif des constructions HLM entre décembre 1993 et décembre 1996, est nettement séparée des autres variables.

Conclusion

La brève analyse présentée ici permet d'insister sur les apports du logiciel POULPE. Ces apports sont particulièrement nets sur deux plans :

- du point de vue méthodologique d'abord : POULPE permet de chiffrer le gain dû au calage sur les marges, technique maintenant largement employée pour redresser les enquêtes ; il permet aussi de juger d'un coup d'oeil la pertinence de l'échantillonnage d'une enquête, par l'examen des design effects ; il facilite enfin la détection de variables 'à problèmes', mal représentées par l'échantillon-maître ou la BSLN ;

- du point de vue du statisticien d'enquête ensuite. POULPE s'avère précieux pour donner des intervalles de confiances sur des fonctions de variables quantitatives, comme le revenu moyen ou les loyers moyens. Si, pour l'enquête Logement, des formules approchées permettent d'approcher correctement la précision d'une large classe d'estimateurs, cela vient sans doute de la taille importante de l'échantillon ; il n'est pas sûr que ce résultat soit valable pour d'autres enquêtes moins lourdes. POULPE permet aussi de juger de la pertinence d'une procédure de redressement, en quantifiant le gain de précision apporté par le calage sur des variables-clés d'une enquête.

La mise à disposition de POULPE, par les facilités qu'elle apporte, a le mérite de rappeler au statisticien d'enquête que la diffusion des résultats d'une enquête devrait systématiquement inclure des indications sur la précision de ces résultats. Les

estimations de précision permettent vis-à-vis des utilisateurs extérieurs de souligner qu'une enquête génère nécessairement des aléas et donc des intervalles de confiance sur les estimations ; elles se révèlent précieuses lorsque certaines évolutions entre deux enquêtes ne sont pas significatives. Les utilisateurs des enquêtes de l'Insee sont en effet plus enclins à s'interroger sur la précision des chiffres qu'il y a quelques années.

Bibliographie indicative

CARON, N. : « Calcul de l'effet de sondage dans le logiciel POULPE », note interne n° 981/F410, 1996.

DEVILLE, J.C. : « Estimation de la précision de données d'enquêtes », document de travail de la Direction des Statistiques Démographiques et Sociales, n° F 9211, Insee, 1992.

DEVILLE, J.C., CARON, N., SAUTORY, O., « Estimation de la précision de données d'enquêtes : document méthodologique sur le logiciel POULPE », document de travail de l'Unité de Méthodologie Statistique, à paraître.

LACROIX, T. : « Pondérations de l'enquête Logement 1992/1993 et révision des pondérations des enquêtes logement 1984 et 1988 », document de travail de la Direction des Statistiques Démographiques et Sociales, n° F 9408, Insee, 1994.

LAFERRERE, A. : « Les ménages et leurs logements », *Insee Première* n° 562, Insee, 1997.

SÄRNDAL, C.E., SWENSSON, B., WRETMAN, J.: *Model assisted Survey Sampling*, Springer-Verlag, 1992

Annexe 1

Variables traitées
Précision et intervalle de confiance à 95 %

Totaux

Variable	Valeur (milliers)	Ecart-type CALMAR	borne inférieure (milliers)	borne supérieure (milliers)
Population des ménages	57785	134773	57521	58049
Nombre de ménages dont la personne de référence a plus de 65 ans	5642	36113	5571	5713
Nombre de propriétaires	12645	39016	12569	12721
Nombre de locataires HLM	3657	21805	3614	3700
Nombre de locataires privés	4449	39447	4372	4526
Nombre de bailleurs privés	1579	32101	1516	1642
Nombre de logements mis en location par des bailleurs privés	2959	124661	2715	3203
Nombre de ménages possédant une résidence secondaire	2051	37555	1977	2125
Nombre de locataires insatisfaits de leur logement	1007	26373	955	1059
Nombre de ménages en loi de 1948	337	17403	303	371
Nombre de ménages dont la personne de référence a entre 20 et 23 ans	640	19759	601	679
Nombre de ménages dont la personne de référence a entre 24 et 27 ans	1264	28703	1208	1320
Nombre de locataires sans bail	797	23147	752	842
Nombre de locataires des deux premiers déciles	2507	37661	2433	2581
Nombre de locataires appartenant aux deux premiers déciles de revenu logés en HLM	1204	25323	1154	1254
Nombre de locataires appartenant aux deux premiers déciles de revenu logés par un bailleur privé	1110	26400	1058	1162
Nombre de ménages à chef indépendant	1092	26143	1041	1143
Nombre de nouveaux ménages	2185	34276	2118	2252
Construction neuve HLM	268	7789	253	283
Nombre d'accédants récents	1658	30222	1599	1717
Nombre de ménages ayant changé de logement dans la même commune depuis 1992	2570	39599	2492	2648

Ratios

Variable	Valeur	Ecart-type CALMAR	borne inférieure	borne supérieure
Revenu mensuel moyen déclaré	13 050	51	12 951	13 149
Prix moyen d'acquisition des logements	669 440	9363	651 089	687 791
Loyer moyen en HLM	1 679	7.6	1 664	1 694
Loyer moyen dans le parc locatif privé	2 517	14.6	2 488	2 546
Nombre moyen de logements mis en location par les bailleurs privés	1.87	7.05E-02	1.73	2.01
Nombre de personnes par ménage	2.48	5.77E-03	2.47	2.49
Part des ménages propriétaires	54.3	0.167	54.0	54.6
Part des ménages insatisfaits	6.0	0.133	5.7	6.3
Part des ménages locataires HLM	15.7	0.093	15.5	15.9
Part des ménages locataires privés	19.1	0.169	18.8	19.4
Part des ménages à chef ouvrier	20.9	0.212	20.5	21.3
Part des ménages bailleurs privés	6.8	0.137	6.5	7.1
Part des locataires logés en HLM	41.1	0.264	40.6	41.6
Part des locataires en locatif privé	50.1	0.340	49.4	50.8
Part des locataires des deux premiers déciles de revenu logés en HLM	25.2	0.480	24.3	26.1
Part des locataires sans bail	9.0	0.258	8.5	9.5
Part des locataires des deux premiers déciles de revenu en difficulté de loyer	29.5	0.820	27.9	31.1
Part des locataires en difficulté de loyer	17.1	0.344	16.4	17.8
Part des ménages sans confort trouvant leurs conditions de logement insuffisantes	12.0	0.400	11.2	12.8
Part des ménages à chef indépendant	4.7	0.112	4.5	4.9
Part des ménages locataires en loi de 1948	1.5	0.074	1.3	1.6

Annexe 2

Application numérique des formules approchées

On cherche à évaluer la précision de trois variables : nombre de locataires du parc privé, part des ménages logés dans le parc locatif privé, part des locataires logés dans le parc privé. Les données tirées de l'enquête sont les suivantes :

Nombre de ménages répondants	29043
Nombre de ménages locataires dans l'échantillon	11096
Nombre de ménages estimé	23 286 000
Nombre de locataires estimé	8 877 000
Nombre de locataires du parc privé estimé	4 449 000
Estimation de la part des ménages logés dans le parc privé	19.1 %
Estimation de la part des locataires logés dans le parc privé	50.1 %

D'où les coefficients de variation approchés suivants :

Quantité	Coefficient de variation approché
nombre de locataires du parc privé	$C_1 = \sqrt{\frac{23286000}{29043}\left(\frac{1-4449000/23286000}{4449000}\right)} = 1.21$
part des ménages logés dans le parc locatif privé	$C_2 = \sqrt{\frac{(1-0.191)}{(29043)(0.191)}} = 1.21$
part des locataires logés dans le parc privé	$C_3 = \sqrt{\frac{(1-0.501)}{(11096)(0.501)}} = 0.95$

On vérifie bien que les coefficients de variation approchés pour le nombre de locataires privés et la part des ménages locataires privés sont les mêmes.