ÉTUDE DES NON-RÉPONSES DANS L'ENQUÊTE EMPLOI

Louis Meuric

Cette note présente les méthodes de redressement et de calage actuellement utilisées dans l'enquête emploi, méthodes qui résolvent en partie les problèmes de biais et de précision dans l'enquête. Elle propose par ailleurs une solution complémentaire fondée sur l'utilisation des données du recensement, afin d'améliorer redressement et calage. Elle donne les étapes de l'étude nécessaire pour valider cette solution alternative, et détaille les conclusions de l'étude des non-réponses.

Problématique : éliminer le biais dû aux non-réponses, améliorer la précision de l'enquête emploi

Présentation des enquêtes emploi

L'enquête emploi est une enquête auprès des ménages, avec comme but d'analyser chaque année en mars la structure de la population active. Emploi, sous-emploi et chômage selon les critères du Bureau International du Travail (BIT), catégorie socio-professionnelle, diplôme, secteur d'activité, temps de travail, salaire et statut correspondant à l'emploi, conditions de recherche d'emploi pour les chômeurs sont autant d'éléments qui permettent de mieux comprendre les mécanismes régissant le marché du travail.

L'échantillon de l'enquête est renouvelé par tiers tous les ans. Depuis 1992, un nouvel échantillon est mis en place progressivement, de sorte qu'en 1994, l'enquête emploi présentera les caractéristiques suivantes :

- 80 000 résidences principales, abritant des ménages ordinaires (au sens du recensement) ;
- 75 000 ménages répondants ;

- 150 000 adultes de 15 ans et plus qui représentent le champ total de l'enquête, soit l'ensemble des individus déclarés comme faisant partie de ces ménages.

De plus, l'échantillon est tiré en dehors de l'échantillon-maître selon les principes suivants ,

- stratification par région et Tranche d'Unité Urbaine (TUU) [21 régions, la Corse étant regroupée avec Paca, 10 TUU, d'où 170 strates]. Les taux de sondage assurent une taille minimale de 5 400 adultes pour toutes les régions, d'où un taux maximal de 1/100^{e} pour le Limousin. Pour les régions qui dépassent cette contrainte de taille, le taux est légèrement inférieur au 1/300^{e}.

- caractère aréolaire : l'échantillon est composé de 3 500 aires (ce sont des zones géographiques délimitées par des frontières nettes et stables). Ces aires comptent 20 logements dans les villes de 100 000 habitants et plus et 40 logements ailleurs.

- tirage à un degré : dans chaque strate, l'échantillon est tiré par sondage aléatoire simple.

Depuis 1992, le tiers sortant est encore interrogé trois fois pour le compte de l'Enquête Trimestrielle sur l'Emploi (ETE). Cette enquête fournit des points conjoncturels sur les grandes catégories de la population active. Elle est réalisée à 88% par téléphone.

Les non-réponses

Celles-ci sont très faibles pour une enquête auprès des ménages (7,1 % en 1992 par exemple). En effet, le fait que les logements soient ramassés les uns sur les autres facilite le travail de relance pour les enquêteurs. Les ETE, réalisées par téléphone, présentent les mêmes taux de non-réponse.

Bien qu'elles soient peu nombreuses, les non-réponses induiraient cependant un biais sur les niveaux de chômage publiés si elles n'étaient redressées de façon appropriée, puisque les non-répondants sont plus souvent au chômage. Ce sont les ETE de 1992 qui le montrent cette fois. En effet, on a tenté d'apparier toutes les résidences principales de l'ETE de juin 1992, par exemple avec les ménages répondant au trimestre précédent, en mars. Sur les 1799 ménages ne répondant pas en juin, 865 avaient pourtant répondu en mars (voir *tableau* page 75).

Ce tableau montre que les personnes qui étaient au chômage en mars ont une probabilité plus importante de ne pas répondre en juin. Il en va de même en décembre, tandis que les non-réponses en septembre sont plutôt de gens ayant un emploi, mais absents de longue durée pour cause de vacances.

On peut craindre que dans l'enquête emploi, les ménages présentent les mêmes comportements de réponse qu'en juin ou décembre. Un redressement des non-réponses adapté est donc nécessaire.

Améliorer la précision des résultats

Du fait de la taille des aires dans le rural (40 logements), on observe un design-effect[1] de 6,6 sur la population des agriculteurs par exemple, de 2 sur le niveau de chômage [2] (voir aussi en annexe quelques écarts-type, calculés en tenant compte des redressements actuels). C'est l'inconvénient du caractère aréolaire de l'échantillon, dont les avantages sont un faible taux de non-réponse et *a priori* une bonne couverture des situations précaires, logements oubliés au RP, locaux transformés en logements, etc... Comme pour toute enquête, il est utile de caler les résultats sur des données exhaustives, fiables, homogènes dans leurs concepts avec ceux de l'enquête, et disponibles trois mois après le début de la collecte, compte tenu des délais de publication des premiers résultats de l'enquête.

Méthode d'estimation employée actuellement

Cette méthode se déroule en deux temps : en premier lieu le redressement des non-réponses, puis indépendemment, le calage de l'échantillon sur la pyramide des âges.

Redressement du biais des non-réponses

Celui-ci est rendu possible du fait que l'on peut dénombrer les résidences principales de l'échantillon, qui représentent le champ de l'enquête, et parmi elles les ménages répondants. On considère actuellement 8 catégories de résidences principales, résultant du croisement du rang d'interrogation et de la TUU. Le tableau ci-dessous donne les taux de non-réponses dans ces 8 catégories en 1992.

	Rural	< 50 000 hab.	> = 50 000 hab.	Agglomération parisienne
Tiers entrant	5,4	7,1	7,6	13,8
Tiers médian et sortant	3,6	5,1	7,4	11,8

(1). Le design-effect est le rapport de la variance d'un estimateur sur un échantillon donné et de la variance du même estimateur sur un échantillon aléatoire simple de même taille.
(2). Journées de méthodologie et statistique de décembre 1991 : L'enquête emploi : échantillon 1992 et années suivantes (N. Roth)

Les non-réponses aux enquêtes emploi (en %)

	Mars 1992	Juin 1992	Sept. 1992	Déc. 1992	Mars 1993	Juin 1993
1 : enquête par téléphone		78,9	78,9	81,1		80,8
2 : enquête par visite		13,1	12,6	12,1		11,8
3 : enquête mixte		0,5	0,4	0,3		0,3
Total	**92,9**	**92,5**	**91,9**	**93,3**	**92,4**	**92,9**
5 : refus 2,3	2,3	1,7	1,7	1,8	2,7	1,9
6 : ménage présent	2,9	3,0	2,9	3,0	3,1	2,6
7 : ALD	1,9	2,8	3,5	1,8	1,8	2,6
Total taux de non-réponses	**7,1**	**7,5**	**8,1**	**6,7**	**7,1**	**7,1**

L'estimateur redressé des non-réponses est un estimateur des valeurs dilatées. Il consiste à diviser le poids de sondage d'un ménage répondant par le taux de réponse de sa catégorie (= 1 - taux de réponse).

Cette méthode peut paraître assez fruste, elle est néanmoins efficace. En effet, si l'on reprend l'étude précédente sur les non-réponses de l'ETE, un modèle logit avec pour explicatives la TUU et les nombres de chômeurs, d'actifs occupés et d'inactifs du ménage au trimestre précédent permet de conclure que les coefficients correspondant à ces trois dernières variables sont nuls. Dans ce modèle, l'activité au trimestre précédent, fortement corrélée avec celle actuelle, n'apporte aucune information supplémentaire une fois pris en compte la TUU et le rang d'interrogation.

En clair, les estimateurs d'emploi et de chômage obtenus par un tel redressement sont correctement corrigés de la non-réponse. Attention, cette étude n'a été bien sûr réalisée que sur les ménages ayant répondu en mars. Il reste à espérer que l'autre moitié des non-répondants, i.e. les récidivistes de la non-réponse, vérifie aussi cette loi.

La question qui se pose maintenant est : un tel redressement suffit-il pour toutes les variables de l'enquête, ou d'autres facteurs interviennent-ils dans le phénomène de la non-réponse? Par ailleurs, cette classification de la TUU est-elle la plus adaptée, ou en existe-t-il une autre plus appropriée à un redressement du même type?

Calage actuel sur la pyramide des âges [1]

En termes de fiabilité, exhaustivité, homogénéité et disponibilité dans un délai de trois mois, on ne possède pour l'instant que la pyramide des âges au moment de l'enquête, par tranche d'âge quinquennal (actualisation du recensement par l'état-civil et par des

(1). Voir "Calage de l'échantillon emploi sur la pyramide des âges", L Meuric, 12 mars 1992, note interne.

Appariement des résidences principales en n avec les répondants en n - 1

	Juin		Septembre		Décembre	
Ménages en n	Total dont répondant en n - 1					
Non-répondants	1 799	soit 7,5%	1 944	soit 8,1%	1 599	soit 6,7%
Répondants	22 192		22 011		22 414	
Ménages de n répondant en n-1						
Non-répondant en n	865	soit 3,9%	796	soit 3,6%	446	soit 2,0%
Répondants	21 225		20 980		21 335	
Réponse en n	Non-réponse	Réponse	Non-réponse	Réponse	Non-réponse	Réponse
Probabilité d'avoir été n-1						
Actif occupé	39,9	47,1	46,7	47,7	47,6	47,4
Chômeur	8,5	5,2	6,9	5,2	11,5	5,9
Inactif	51,6	47,8	46,4	47,1	40,9	46,7
Total	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**

hypothèses de migration). La méthode de calage utilisée est le Raking Ratio Généralisé (RRG), développée par C.-E. Sarndal et J.-C. Deville[1] et programmée par O. Sautory dans la macro SAS CALMAR.

Le Raking Ratio simple permet déjà de caler sur les effectifs marginaux de deux variables qualitatives ou plus sans avoir à caler sur leurs effectifs croisés. Le RRG quant à lui, permet également d'introduire des variables quantitatives, résolvant ainsi le problème de la cohérence des statistiques individuelles et des statistiques de ménages dans l'enquête emploi.

En effet, dans l'ancienne série, le calage sur la pyramide des âges conduisait à des poids différents pour les membres d'un même ménage. Ce calage consistait à calculer le rapport entre l'effectif d'une tranche d'âge après redressement des non-réponses et celui de la source officielle, puis à multiplier le poids de tous ses membres par ce rapport.

(1). "Estimateurs par calage et techniques de ratissage généralisé dans les enquêtes par sondage", C.E. Sarndal et J.C. Deville, note interne.

Quel poids alors retenir pour le ménage, comment concilier les statistiques sur les hommes et les femmes vivant en couple par exemple?

Pour que les membres d'un ménage aient le même poids et que l'on puisse cependant caler sur des données individuelles exhaustives, il suffit de considérer pour chaque ménage les variables quantitatives que sont les nombres d'hommes et de femmes du ménage appartenant à telle ou telle tranche d'âge, et d'en ajuster les totaux calculés sur les ménages de l'échantillon.

Malgré ce calage, l'écart-type sur les effectifs des agriculteurs reste important (40 000 pour une population de 1 043 000 en 1992). Il faut donc trouver d'autres sources exhaustives que la pyramide des âges, mieux corrélées avec nos variables d'intérêt, le recensement de 1990 par exemple. La solution proposée maintenant repose elle aussi sur l'utilisation du RRG, tant pour le redressement des non-réponses que pour les calages.

Solution complémentaire : affiner le redressement des non-réponses, caler aussi sur le RP, articuler tout cela

Propositions

Elles consistent à exploiter au maximum les caractéristiques RP des résidences principales présentes dans les aires au moment du recensement, et les caractéristiques au moment de l'enquête des nouvelles résidences principales, connues même si leurs occupants ne répondent pas. Dans ce but, on doit définir 3 catégories de logements :

- Les résidences principales lors du RP90 qui ne le sont plus au moment de l'enquête. Il peut s'agir d'une destruction ou d'une transformation en résidence secondaire, en logement vacant ou occasionnel). Ces logements sont exclus du champ de l'enquête, mais vont pourtant servir ;

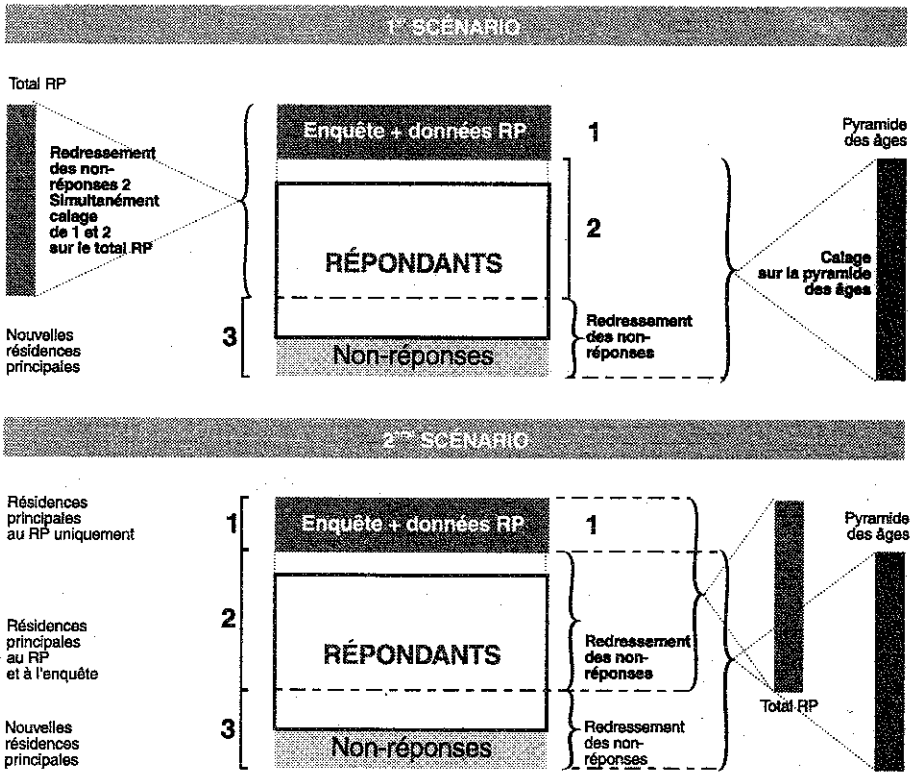
- Les résidences principales lors du RP90 qui le sont toujours à la date de l'enquête:
- Les nouvelles résidences principales à la date de l'enquête :
  - logements neufs : pas d'information au RP, mais uniquement à l'enquête (type d'immeuble, nombre de logements, année d'achèvement) ;
  - anciennes résidences secondaires, anciens logements vacants ou occasionnels: on dispose alors d'informations sur le logement au RP (type d'immeuble, nombre de logements, année d'achèvement, nombre de pièces), mais pas sur le ménage lors du RP. Sur cette faible population, se restreindre aux variables de l'enquête ne fait perdre que l'information sur le nombre de pièces.

Pour les catégories 1 et 2 par contre, qui sont les résidences principales au moment du RP, toutes sortes d'informations sont disponibles sur le ménage et le chef de ménage. Lorsque l'information complète fait cependant défaut (problème d'appariement avec le RP), on peut imputer les caractéristiques actuelles ou des caractéristiques aléatoires, la question n'est pas encore tranchée.

Cette information permet d'une part d'affiner le redressement des non-réponses pour la catégorie 2, d'autre part d'améliorer la précision des résultats en calant les résidences principales RP de l'échantillon (catégories 1 + 2) sur les résultats exhaustifs du RP, d'où l'intérêt de la catégorie 1. Il va de soi que les gains en biais et en précision ainsi acquis diminueront avec le temps, au fur et à mesure que s'atténuera la corrélation des variables auxiliaires avec le phénomène de non-réponses et avec les variables d'intérêt.

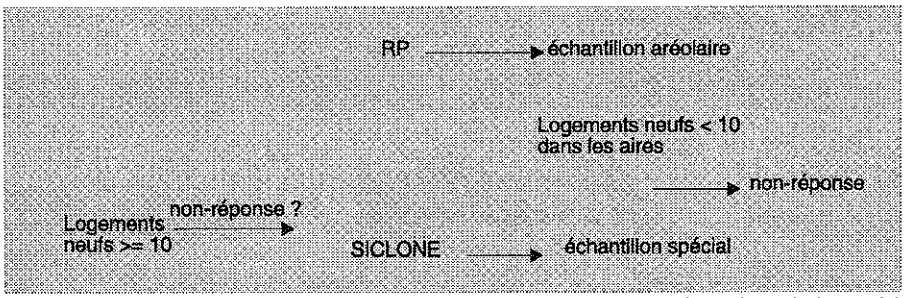
Dans tous les cas, le tiers sortant est ensuite enquêté pour le compte de l'enquête trimestrielle. Par souci d'homogénéité, ce tiers est donc traité à part. Le tiers entrant présentant davantage de non-répondants, il le sera également. En clair, les trois tiers de l'enquête seront redressés des non-réponses et calés séparément.

Le tout est maintenant d'articuler redressement et calage, deux scenarii étant possibles.



L'avantage du deuxième scénario est qu'il opère les deux calages en même temps, le calage sur la pyramide des âges ne vient donc pas détruire celui sur le RP. De plus, il permet paradoxalement de mettre en œuvre un modèle plus complexe de non-réponse, tenant compte d'un suivi des logements neufs dans l'enquête éventuellement défectueux.

En effet, les Services statistiques des établissements régionaux[1] reçoivent pour consigne de constituer des aires géographiques aussi représentatives que possible de leur commune ou de leur district. Ainsi, les aires doivent éventuellement contenir des terrains à bâtir, afin de bien prendre en compte la construction neuve. Mais avec des immeubles neufs de 6, 12 ou 20 étages par exemple, elles deviendraient rapidement bien trop importantes, d'où un effet grappe désastreux et une charge de travail déséquilibrée. Les enquêteurs ont donc pour consigne d'exclure tout immeuble ou lotissement construit sur permis de 10 logements ou plus, ceux-ci faisant chaque année l'objet d'échantillonnages spéciaux à partir des fichiers des logements neufs SICLONE. Mais ces fichiers n'étant pas parfaitement exhaustifs, on court le risque de sous-estimer les logements neufs, souvent occupés par des ménages jeunes et de taille réduite, d'où la nécessité du modèle suivant:



Le calage simultané sur le RP et sur la pyramide des âges devrait maintenir les poids des anciennes résidences principales tout en surpondérant éventuellement les logements neufs d'un effet non-réponse. Une telle méthode n'est applicable que si la notion de résidence principale est bien verrouillée à l'enquête. On n'exclut donc pas de devoir mettre en œuvre le premier scénario.

Études à réaliser pour valider ces propositions

Elles consistent à :
- étudier les non-réponses dans l'enquête emploi, tant pour les résidences principales de catégorie 2 que pour celles de catégorie 3, afin d'affiner les redressements ;
- déterminer quelles variables auxiliaires du RP doivent figurer dans le calage: ce sont les variables les mieux corrélées avec les variables d'intérêt que sont l'activité (chômeur, inactif ou en emploi) et la catégorie socioprofessionnelle. Attention, on

(1) Les services statistiques sont chargés de faires les enquêtes.

doit tenir compte du fait que l'échantillon est déjà stratifié par région et tranche d'unité urbaine, et que l'on cale toujours sur la pyramide des âges ;
- enfin, examiner quelle est la meilleure articulation du ou des calages avec le redressement des non-réponses.

Seules les conclusions sur les non-réponses sont exposées ci-après.

Étude des non-réponses

Méthode

L'objectif est d'assurer le meilleur redressement des non-réponses possible à l'enquête emploi, ce qui implique:

- d'étudier ces non-réponses toutes catégories confondues, qu'il s'agisse de refus, d'absents de longue durée ou que l'enquête soit impossible à réaliser. Cette étude se fera au moyen d'un modèle logit ;

- que les variables explicatives rentrées dans ce modèle devront être les plus discriminantes possible: la procédure de test pas à pas de chacune des variables (nous verrons lesquelles tout à l'heure) permet d'atteindre ce but ;

- que ces variables auxiliaires soient bien corrélées avec nos variables d'intérêt (activité, CS). De plus, deux catégories de variables auxiliaires se distinguent: celles du logement, stables, et celles concernant le ménage au moment du RP, dont la corrélation avec le phénomène de non-réponses diminue avec le temps. **On privilégiera donc les caractéristiques du logement** s'il y a un choix à faire et l'on étudiera la significativité des caractéristiques du ménage conditionnellement aux premières.

Mise en oeuvre de modèles logit sur 2 populations : anciennes et nouvelles résidences principales (catégories 2 et 3)

Préliminaires

Catégorie 2 : parmi toutes les caractéristiques RP des logements, des ménages ou des chefs de ménage, on a choisi les variables *a priori* les plus susceptibles d'influer sur les comportements de non-réponse.

Catégorie 3 : toutes les variables disponibles à l'enquête ont été retenues. Leur liste figure ci-dessous.

Des tabulations croisées avec la variable RÉPONSE ci-dessous et des tests du chi-deux ont ensuite permis d'en éliminer quelques-unes comme trop peu corrélées avec la non-réponse (situation particulière d'emploi du chef de ménage, confort du logement), et aussi d'effectuer des regroupements de modalités pour les variables âge du chef de ménage et type de logement (voir plus loin), en fonction des taux de non-réponse dans chacune de leurs modalités.

Enfin, certaines variables quantitatives ont été tronquées par le haut ou regroupées afin que chaque modalité compte suffisamment de non-répondants. Il s'agissait des nombres de personnes, d'adultes, d'enfants, d'actifs occupés du ménage et du nombre de logements de l'immeuble.

	Anciennes résidences principales	**Nouvelles résidences principales**
Population totale	47 060	6 096
Non répondants	3 476	733
Variables explicatives disponibles	**RP**	**Enquête**
	Logement	TUU
	TUU, année d'achèvement	Année d'achèvement
	Type de logement	Type de logement.
	Nombre de logements	Nombre de logements
	Nombre de pièces	
	Statut d'occupation confort	
	Ménage	
	Nombres de personnes	
	d'adultes,d'enfants,d'actifs	
	Chef de ménage	
	Sexe,âge,nationalité	
	Activité,statut	
	Situation particulière d'emploi	
	Position professionnelle	

Mise en œuvre de la procédure logistic

Dans les deux cas, on a procédé de la même manière :

- *variable dépendante* : RÉPONSE = 1 si le ménage répond
  0 sinon: on ne distingue pas le type de non-réponse
- *variables explicatives:* celles décrites ci-dessus. L'habitude est de les déclarer sous forme de variables "dummies" ou dichotomiques, ou encore appelées indicatrices, de la façon suivante: si une variable qualitative, le nombre NP de personnes du ménage par exemple, compte K modalités, on définit:

$$NP1 = 1 \text{ si } NP = 1$$
$$0 \text{ sinon}$$
$$NP2 = 1 \text{ si } NP = 2$$
$$0 \text{ sinon}$$
$$\vdots$$
$$NPK1 = 1 \text{ si } NP = K - 1$$
$$0 \text{ sinon}$$

On s'arrête à K-1 car si l'on introduisait dans le modèle l'indicatrice correspondant à la dernière modalité, la somme des indicatrices de 1 à K donnant 1, il y aurait colinéarité entre les variables explicatives. On retire donc usuellement la dernière indicatrice. Les coefficients des autres indicatrices s'interprètent alors comme des écarts par rapport à la dernière modalité, dont l'effet sur la non-réponse est arbitrairement fixé à zéro.

Mais ce système d'indicatrices ne convient pas à notre but initial: tester la corrélation des variables qualitatives et surtout tester les regroupements de leurs modalités. On a préféré retenir les indicatrices définies ci-dessous adaptées à cet objectif:

$$NP1 = 1 \text{ si } NP \Leftarrow 1$$
$$0 \text{ sinon}$$
$$NP2 = 1 \text{ si } NP\,2 \Leftarrow 2$$
$$0 \text{ sinon}$$
$$\vdots$$
$$NPK1 = 1 \text{ si } NP = K - 1$$
$$0 \text{ sinon}$$

En effet, tester la nullité du coefficient de NP1 par exemple équivaut à tester qu'on puisse regrouper les modalités 1 et 2, toutes choses égales par ailleurs. De sorte qu'alors, la procédure stepwise de test automatique de nullité de tous les coefficients peut être mise en oeuvre. Les indicatrices restantes mettent alors en évidence les seules frontières vraiment pertinentes, et par complémentarité, les regroupements de modalités appropriés : si NP2 demeure, cela signifie que le fait que le ménage compte jusqu'à 2 personnes ou qu'il en compte davantage est discriminant quant à la non-réponse.

Toujours concernant les variables quantitatives, on constate souvent des taux de non-réponse croissants ou décroissants : il peut donc être intéressant de tester que la non-réponse est directement proportionnelle à la variable numérique ou que certaines modalités ont des effets spécifiques. Pour le nombre de personnes par exemple, on a ainsi mis en concurrence les K-1 variables dichotomiques NP1---NPK-1 et la variable quantitative NP.

Qu'en est-il des variables a priori non ordonnées?

En fait, il n'est pas nécessaire de disposer d'un ordre naturel pour chaque variable qualitative; il suffit d'une batterie de tests de regroupements de modalités ad hoc, autorisant chaque catégorie à ne se regrouper qu'avec 2 catégories voisines au maximum, et à interdire tout autre regroupement. Ainsi, si l'on considère le type d'immeuble au recensement :

Type d'immeuble	Non-répondants	Taux de non-réponse
(1) maison individuelle	1 398	5,47
(2) immeuble	1 958	10,31
(3) foyer	18	6,04
(4) ferme	39	2,96
(5) hôtel	15	20,83
(6) habitation de fortune	3	16,67
(7) pièce indépendante	28	10,65
(8) logement non à usage d'habitation	17	3,22

On a ordonné cette variable selon les quatre modalités suivantes:

A: (4),(8)
B: (1),(3)
C : (2),(6),(7)
D : (5)

La nature des modalités considérées permet donc souvent de restreindre la batterie de tests comme si l'on avait affaire à une variable ordonnée. A défaut, on peut aussi trier les modalités par taux de non-réponse croissant.

Arrivé à ce stade, on connaît les critères les plus discriminants dans l'explication des non-réponses, ordonnés d'ailleurs par la procédure stepwise par ordre décroissant d'importance. Pour les anciennes résidences principales, il peut alors être intéressant de croiser les plus importants, qui étaient le nombre de pièces, le fait d'habiter à Paris ou ailleurs, le fait d'habiter dans une maison ou un immeuble (catégories A, B versus catégories C, D). Cela n'a cependant rien donné de concluant.

Conclusions

Un test du rapport du maximum de vraisemblance montre que la TUU et le rang d'interrogation ne suffisent pas à expliquer la non-réponse, qu'il s'agisse de l'une ou l'autre catégorie de logement étudiées.

a) Les nouvelles résidences principales

Apparaissent comme largement discriminants, et par ordre :
- la TUU: les Parisiens (intra-muros) répondent moins souvent ;
- le type d'immeuble : les ménages vivant dans des immeubles répondent moins souvent, peut-être du fait des digicodes ou autres barrières à l'entrée ;
- l'année d'achèvement de l'immeuble (avant 1982 ou après).

b) Les anciennes résidences principales

On a fait tourner deux modèles logit : le premier sur les seules caractéristiques du logement , le second y compris sur les caractéristiques du ménage et du chef de ménage. Dans les deux modèles, ce sont les mêmes caractéristiques du logement qui expliquent la non-réponse : il n'y a donc pas de choix à faire entre ces dernières et les caractéristiques du ménage.

Par ailleurs, les caractéristiques du ménage sont significatives conditionnellement à celles du logement (test du rapport du maximum de vraisemblance). On doit bien prendre en compte les caractéristiques du ménage, même si l'on craint que le gain pour correction des non-réponses qu'elles apporteront diminuera davantage avec le temps.

Apparaissent comme largement discriminants, et par ordre :
- le nombre de pièces (variable quantitative) : effet décroissant ;
- la TUU: les Parisiens (intra-muros) répondent moins souvent ;

- le nombre de personnes (variable quantitative) : effet décroissant ;
- le type d'immeuble (immeuble ou maison individuelle) ;
- le rang d'interrogation: en deuxième enquête, les gens répondent mieux ;
- l'âge du chef de ménage.

D'autres variables sont également discriminantes, en ce sens que le test du stepwise à 5% les a retenues, mais leur apport est plus faible (voir en annexe) : si elles ne se combinent pas naturellement avec les variables ci-dessus, on ne les retiendra pas. Ainsi, on gardera les différentes catégories de tranche d'unité urbaine, complémentaires du lieu de résidence, mais on excluera l'année d'achèvement de l'immeuble, le nombre d'actifs du ménage, le nombre de logements de l'immeuble et le type d'activité du chef de ménage. En sorte qu'après un nouveau test sans ces variables, on obtient les catégories des variables ci-dessus devant être utilisées pour le redressement des non-réponses. Elles figurent en dernière page de l'annexe.

Il reste à définir les variables RP les plus pertinentes pour le calage, et à étudier comment articuler calage et redressement des non-réponses, selon la méthode exposée en page 78.

ANNEXE

Tableau 5 bis (suite) : *Les erreurs aléatoires de l'enquête*

Variable	Valeur de la variable et intervalle de confiance à 95 %		
	Hommes	Femmes	Les deux sexes
POPULATION ACTIVE OCCUPEE (en nombre)			
Actifs occupés au sens du BIT	12 784 + ou - 76	9 548 + ou - 91	22 330 + ou - 135
Salariés	10 630 + ou - 92	8 381 + ou - 91	19 011 + ou - 149
Non-salariés	2 154 + ou - 76	1 165 + ou - 52	3 319 + ou - 117
1. Agriculteurs exploitants	650 + ou - 50	393 + ou - 34	1 043 + ou - 79
2. Artisans, commerçants, chefs d'entreprises	1 173 + ou - 47	582 + ou - 34	1 755 + ou - 71
3. Cadres et professions intellectuelles supérieurs	1 856 + ou - 81	849 + ou - 46	2 704 + ou - 115
4. Professions intermédiaires	2 524 + ou - 66	1 961 + ou - 58	4 485 + ou - 101
5. Employés	1 369 + ou - 61	4 532 + ou - 81	5 901 + ou - 110
6. Ouvriers	4 963 + ou - 108	1 228 + ou - 53	6 190 + ou - 140
01. Agriculture			1 311 + ou - 89
02. Industries agricoles et alimentaires			529 + ou - 39
03. Energie			240 + ou - 26
04. Industries des biens intermédiaires			1 171 + ou - 61
05. Industries des biens d'équipement			1 482 + ou - 63
06. Industries des biens de consommation			1 203 + ou - 53
02 à 06. Industrie (en milliers)			4 724 + ou - 132
07. Bâtiment, génie civil et agricole			1 369 + ou - 56
08. Commerce (en milliers)			2 607 + ou - 70
09. Transports et télécommunications			1 365 + ou - 56
10. Services marchands			5 228 + ou - 101
11 à 13. Institutions financières			711 + ou - 39
14. Services non marchands			4 709 + ou - 115
08 à 14. Tertiaire			14 619 + ou - 164
STAGIAIRES			
Nombre total des stagiaires actifs occupés	674 + ou - 33	616 + ou - 29	1 290 + ou - 47

The SAS System 16:09 Wednesday, October 6, 1993

The LOGISTIC Procedure Nouvelles résidences principales

No (additional) variables met the 0.05 significance level for entry into the model.

Summary of Stepwise Procedure

Step	Variable Entered	Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi-Square
1	TU9 Province/Paris		1	109.4	.	0.0001
2	TH4 Maison/Immeuble		2	50.2704	.	0.0001
3	A5 Année d'achèvement: ≤ 91		3	27.6241	.	0.0001
4	A4 ≤ 74		4	6.7453	.	0.0094
5	TU1 Canton entièrement rural		5	5.0212	.	0.0250
6	LO1 Un logement dans l'immeuble		6	4.2998	.	0.0381
7	TU6 Moins de 100 000 habitants		7	4.7135	.	0.0299

Analysis of Maximum Likelihood Estimates -2 log vraisemblance = 4460

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.6703	0.1074	241.7339	0.0001	.	0.188
TU1	1	-0.5530	0.2166	6.5202	0.0107	-0.080296	0.575
TU6	1	0.2215	0.1021	4.7037	0.0301	0.061050	1.248
TU9	1	-0.7098	0.1048	45.8973	0.0001	-0.143700	0.492
LO1	1	-0.2599	0.1134	5.2530	0.0219	-0.070741	0.771
A4	1	-0.2873	0.1239	5.3752	0.0204	-0.077724	0.750
A5	1	0.7410	0.1398	28.0768	0.0001	0.188979	2.098
TH4	1	-0.5016	0.1241	16.3464	0.0001	-0.137610	0.606

Association of Predicted Probabilities and Observed Responses

Concordant	= 62.3%	Somers' D	= 0.314
Discordant	= 31.0%	Gamma	= 0.336
Tied	= 6.7%	Tau-a	= 0.060
(4468368 pairs)		c	= 0.657

The SAS System 19:26 Tuesday, November 9, 1993 1

The LOGISTIC Procedure Anciennes résidences principales: rang d'interrogation x TUU

Data Set: WORK.B
Response Variable: REP
Response Levels: 2
Number of Observations: 47060
Link Function: Logit

Response Profile

Ordered Value	REP	Count
1	0	3476
2	1	43584

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	24804.408	24418.877	.
SC	24813.167	24488.950	.
-2 LOG L	24802.408	24402.877	399.531 with 7 DF (p=0.0001)
Score	.	.	438.442 with 7 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-2.0285	0.0529	1472.7365	0.0001	.	0.132
1 interrogation RAN total	1	0.2029	0.0724	7.8594	0.0051	0.055934	1.225
TU1 Rural	1	-0.2293	0.0829	7.6470	0.0057	-0.055600	0.795
TU2 < 50000	1	-0.1440	0.0718	4.0194	0.0450	-0.039707	0.866
TU3 Hors Paris	1	-0.6118	0.0688	78.9819	0.0001	-0.119355	0.542
TU4 Rural	1	-0.0909	0.1125	0.6532	0.4190	-0.016974	0.913
TU5 < 50000	1	0.0160	0.0970	0.0271	0.8693	0.003811	1.016
TU6 Hors Paris	1	0.0269	0.0937	0.0823	0.7742	0.007328	1.027

Association of Predicted Probabilities and Observed Responses

Concordant	= 53.0%	Somers' D	= 0.190
Discordant	= 34.0%	Gamma	= 0.218
Tied	= 13.0%	Tau-a	= 0.026
(151497984 pairs)		c	= 0.595

The SAS System 13:55 Saturday, November 6, 1993 6

The LOGISTIC Procedure Anciennes résidences principales : caractéristiques du logement 1ère tentative.

o (additional) variables met the 0.05 significance level for entry into the model.

Summary of Stepwise Procedure

Step	Variable Entered	Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi-Square
1	NPCE		1	559.1	.	0.0001
2	TU9		2	167.1	.	0.0001
3	TH2		3	68.3623	.	0.0001
4	RANG		4	36.2065	.	0.0001
5	A4		5	30.4121	.	0.0001
6	TH1		6	23.7025	.	0.0001
7	A1		7	9.8488	.	0.0017
8	TU1		8	9.9446	.	0.0016
9	L04		9	6.1488	.	0.0131
10	N6		10	4.3509	.	0.0370
11	TU5		11	3.9925	.	0.0457
12	TU3		12	8.0221	.	0.0046
13	TU6		13	6.9717	.	0.0083

- 2 log vraisemblance = 23904

Analysis of Maximum Likelihood Estimates

Variable		Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT		-1.3631	0.1380	97.6176	0.0001		0.256
RANG	4ème ou 2ème interrogation	0.1108	0.0179	38.3211	0.0001	0.061064	1.117
NPCE	Nombre de pièces	-0.2278	0.0169	180.7697	0.0001	-0.146967	0.796
N6	Moins de 6 pièces	-0.2175	0.1051	4.2826	0.0385	-0.024273	0.805
TU1	TU1: commune rurale	-0.2912	0.0909	10.2583	0.0014	-0.044502	0.747
TU3	5000 hab	-0.2185	0.0760	8.2649	0.0040	-0.056343	1.244
TU5	5000 à 20 000	-0.3840	0.0895	18.4069	0.0001	-0.104879	0.681
TU6	20 000 à 100 000	0.1869	0.0709	6.9546	0.0084	0.051530	1.206
TU9	Ouvrier/Paris	-0.3982	0.0685	67.3187	0.0001	-0.077693	0.672
A1	Année 1993 et 1994	0.1585	0.0453	12.2607	0.0005	0.037366	1.172
A4	Année 1995 et 1996	0.2131	0.0658	21.6425	0.0001	0.052870	1.238
TH1	Travaux manuels	-0.6560	0.1411	21.6286	0.0001	-0.070211	0.519
TH2	Autres travaux	-0.2674	0.0492	29.5967	0.0001	-0.072540	0.765
L04	Moins de 14 logements	-0.1184	0.0492	5.8011	0.0160	-0.025297	0.888

Association of Predicted Probabilities and Observed Responses

Concordant	= 63.5%	Somers' D	= 0.294
Discordant	= 34.1%	Gamma	= 0.302
Tied	= 2.5%	Tau-a	= 0.040
(151497984 pairs)		c	= 0.647

The LOGISTIC Procedure

Summary of Stepwise Procedure

Step	Variable Entered	Variable Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi-Square
1	NPCE		1	559.1	.	0.0001
2	TU9		2	167.1	.	0.0001
3	NP		3	156.8	.	0.0001
4	TH2		4	71.3659	.	0.0001
5	RANG		5	36.9279	.	0.0001
6	AG6		6	34.4734	.	0.0001
7	TH1		7	17.8541	.	0.0001
8	AC5		8	16.7764	.	0.0003
9	A4		9	13.4115	.	0.0005
10	AG3		10	12.0623	.	0.0003
11	NP5		11	13.0193	.	0.0010
12	AG1		12	10.8631	.	0.0045
13	TU1		13	8.0672	.	0.0070
14	A1		14	7.2641	.	0.0098
15	NAC3		15	6.6701	.	0.0101
16	AC2		16	6.6172	.	0.0176
17	N6		17	5.6397	.	0.0227
18	TU5		18	5.1912	.	0.0251
19	TU3		19	7.8558	.	0.0272
20	TU6		20	4.8773	.	0.0306
21	L04		21	4.6755	.	

Analysis of Maximum Likelihood Estimates

Variable		Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT		0.0787	0.2323	0.1149	0.7346		1.082
RANG		0.1115	0.0179	38.6465	0.0001	0.061497	1.118
		-0.1086	0.0196	30.7476	0.0001	-0.070069	0.897
HPCE		-0.2472	0.1057	5.4730	0.0193	-0.027504	0.781
N6		-0.2858	0.0911	9.8338	0.0017	-0.043682	0.751
TU1		0.2152	0.0762	7.9658	0.0048	0.055478	1.240
TU3		-0.3702	0.0898	16.9951	0.0001	-0.101089	0.691
TU5		0.1642	0.0711	5.3315	0.0209	0.045269	1.178
TU6		-0.4463	0.0491	82.8002	0.0001	-0.087083	0.640
TU9		0.1237	0.0454	7.4266	0.0064	0.029165	1.132
A1		0.1522	0.0471	10.4348	0.0012	0.037755	1.164
A4		-0.5823	0.1414	16.9505	0.0001	-0.062320	0.559
TH1		-0.2792	0.0500	31.2320	0.0001	-0.075726	0.756
TH2		-0.1066	0.0493	4.6733	0.0306	-0.022780	0.899
LA4		-0.1919	0.0593	10.4774	0.0012	-0.034414	0.825
AG1	Age : <30	0.2509	0.0466	28.9596	0.0001	0.069135	1.285
AG3	<50	-0.3649	0.0635	33.0352	0.0001	-0.050502	0.694
AG6	<80	-0.4822	0.1437	11.2645	0.0008	-0.046138	0.617
NP5	6 pieces	-0.2587	0.0210	152.3144	0.0001	-0.194551	0.772
HP	Nombre de pièces	0.1555	0.0575	7.3130	0.0068	-0.028401	0.856
AC2	Etudiant,	-0.1871	0.0577	10.4996	0.0012	-0.029034	0.829
AC5		-0.2481	0.0939	6.9805	0.0082	-0.031972	0.780
NAC3							

The SAS System

The LOGISTIC Procedure

Ancienne résidences principales : modèle final

No (additional) variables met the 0.05 significance level for entry into the model.

Summary of Stepwise Procedure

Step	Variable Entered	Variable Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi-Square
1	NPCE		1	559.1	.	0.0001
2	TU9		2	167.1	.	0.0001
3	NP		3	156.8	.	0.0001
4	TH2		4	71.3659	.	0.0001
5	RANG		5	36.9279	.	0.0001
6	AG6		6	34.4734	.	0.0001
7	TH1		7	17.0541	.	0.0002
8	A4		8	14.1597	.	0.0002
9	AG3		9	11.3468	.	0.0008
10	AG2		10	13.7086	.	0.0002
11	TU1		11	7.8401	.	0.0051
12	A1		12	7.9173	.	0.0049
13	AG1		13	4.3246	.	0.0376
14	TU5		14	4.2957	.	0.0382
15	TU3		15	8.3180	.	0.0039
16	TU6		16	5.3627	.	0.0206

20:50 Tuesday, November 16, 1993 8

The SAS System

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable		Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT		-1.1351	0.0957	140.6657	0.0001		0.321
RANG		0.1101	0.0179	37.6867	0.0001	0.060678	1.116
NPCE		-0.1172	0.0192	37.2006	0.0001	-0.075621	0.889
TU1		-0.2856	0.0911	9.8246	0.0017	-0.043647	0.752
TU3		-0.2225	0.0762	8.5329	0.0035	0.057373	1.249
TU5		-0.3718	0.0897	17.1838	0.0001	-0.101549	0.689
TU6		0.1642	0.0710	5.3526	0.0207	0.065253	1.178
TU9		-0.4587	0.0482	90.6060	0.0001	-0.089494	0.632
TH1		-0.5751	0.1412	16.5779	0.0001	-0.061550	0.563
TH2		-0.3196	0.0672	45.9242	0.0001	-0.084688	0.726
A1		0.1199	0.0448	7.1531	0.0075	0.028271	1.127
A4		0.1486	0.0466	10.1812	0.0014	0.036875	1.160
AG1	Age < 30	-0.1381	0.0653	4.4779	0.0343	-0.024777	0.871
AG2	< 40	-0.1451	0.0599	5.8696	0.0154	-0.031443	0.865
AG3	< 50	0.2734	0.0547	24.9895	0.0001	0.075357	1.314
AG4	< 60	0.3795	0.0633	35.9845	0.0001	0.052525	1.684
NP		-0.2135	0.0186	131.1880	0.0001	-0.160603	0.808