

APPLICATION D'UNE MÉTHODOLOGIE DES « PETITS DOMAINES » Estimation de l'isolement relationnel dans 3 zones urbaines sensibles

Jean-Louis PAN KÉ SHON^(*), Henri VIVIER^(**)

^(*) Insee, Division des études territoriales

^(**) Insee, Direction régionale de Bretagne

Ce document est la prolongation d'une première étude sur la non-inscription électorale bénéficiant des réflexions du groupe de travail réunissant outre les auteurs de cet article : Valérie Albouy de la Division des études sociales, Insee, Pascal Ardilly de l'Unité de Méthodologie Statistique, Insee, Claude Gissot de la Drees. Bien sûr, ils ne sont pas redevable des éventuelles erreurs du document. Les auteurs les remercient chaleureusement de leurs apports précieux.

Introduction

La croissance des besoins d'informations économiques et sociales aux divers niveaux géographiques infra nationaux va de pair d'une part avec l'importance grandissante de la décentralisation et d'autre part avec l'expansion de l'Europe et des besoins d'évaluations comparatives. Dans un cas comme dans l'autre, il n'est pas difficile de déduire que ces besoins ne devraient pas se tarir au cours des prochaines années ! Les Directions régionales de l'Insee sont ainsi fréquemment sollicitées par divers acteurs sur des sujets pour lesquelles elles n'ont souvent que des données représentatives au niveau national. En l'absence de toutes données représentatives au niveau infra national, il n'est possible de se référer alors qu'à une moyenne nationale qui est attribuée avec plus ou moins de précautions au domaine réduit d'étude. Obtenir des estimations « raisonnablement fiables » à un niveau infra national, c'est-à-dire relativement plus fiables que l'estimation globale nationale, est l'ambition modeste et tout l'intérêt de la méthodologie des petits domaines. En effet, l'estimation petit domaine ne demeure qu'un succédané d'estimation directe et ne s'y substitue que par défaut.

Par ailleurs, l'étude a été volontairement placée dans des conditions inconfortables, l'échantillon de l'enquête Permanente Conditions de Vie est relativement faible (11 500 individus) et le domaine à estimer est réduit (près de 19 000 individus). D'une part afin d'évaluer les possibles et d'autre part à cause du seul point de comparaison avec un estimateur direct sur la zone à étudier autorisé par l'extension régionale de l'enquête de Bretagne.

Ce travail a pour double but d'estimer l'isolement relationnel dans une zone urbaine sensible (Zus) ou un groupe de Zus et de mettre au point une méthode d'estimation infra régionale à partir de données n'ayant qu'une représentativité nationale et applicable à diverses variables qualitatives. Le document chemine à travers les différentes étapes nous ayant permis de réaliser les premières estimations jusqu'aux contrôles de ces résultats. L'optique retenue est de s'inscrire dans une démarche d'un chargé d'études face à l'expertise d'une demande nécessitant l'utilisation de cette méthodologie.

1. Le support d'étude

La sociabilité des individus est souvent perçue au travers de ses aspects ludiques, sorties, réceptions, loisirs partagés en commun héritant en cela des premières études statistiques françaises (Lemel, Paradeise, 1967, Forsé, 1981, Choquet, 1982) ou monographiques (Bozon, 1984). Mais elle ne se réduit pas à cette seule dimension hédoniste même prise sous l'angle des inégalités. Elle est d'une part un élément essentiel de la construction de soi dans les rapports aux autres, une socialisation dont les effets perdurent au-delà de la jeunesse. D'autre part la sociabilité est un facteur d'économie informelle, de redistribution au sein de la famille (voir notamment Marpsat, 1991, Wolff, 2000a, 2000b). Elle représente également un vecteur de circulation d'informations et d'interventions diverses, de parrainage déjà étudiée sous l'angle de l'accès à l'emploi (Degenne et alii, 1991, Forsé, 1994, 2000, Giret et alii, 1996, Marry, 1983). En bref, un capital social dans l'acception française du terme (Bourdieu, 1980a, 1980b) qui n'est pas indépendant de l'échelon occupé sur l'échelle sociale mais qui est irréductible à celui-ci.

A l'inverse, l'isolement relationnel dévoile un certain « déficit social » des individus et peut se définir conventionnellement comme un indicateur signalant les personnes « pauvres » en contacts interpersonnels. Combiné avec des situations précaires, l'isolement peut entraîner des états de rupture : « ...l'absence de participation à toute activité productive et l'isolement relationnel conjuguent leurs effets négatifs pour produire l'exclusion ou plutôt(...) la désaffiliation. La vulnérabilité sociale est une zone intermédiaire instable, qui conjugue la précarité du travail et la fragilité des supports de proximité. » (Castel, 1995). Ainsi, l'isolement relationnel renvoie à l'usure du lien social entre individu fragilisé et société rendue incapable de préserver le contrat social.

Sur le territoire national, les zones urbaines sensibles (Zus) concentrent des populations défavorisées et l'estimation quantifiée de l'isolement au sein de ces quartiers devrait amener à produire un indicateur de fragilisation complétant les portraits locaux des populations en Zus. En l'absence d'estimation directe possible, l'estimation par méthode des « petits domaines » se révèle précieuse.

2. L'isolement relationnel, un concept hétérogène

L'isolement relationnel traduit donc une « certaine » pauvreté en contacts interpersonnels et entraîne la prise en compte des problèmes liés à la définition de la pauvreté, notamment pour les aspects conventionnel et normatif (voir notamment Économie et statistique n° 308-310, 1997). Il est nécessaire de souligner la création induite de norme par l'utilisation de cet indicateur puisque implicitement être isolé relationnellement revient à désigner une faiblesse constitutionnelle de la personne, en oubliant pour une part les personnes s'isolant volontairement. Il reste qu'on peut émettre l'hypothèse raisonnable que la grande majorité des isolés subit cet isolement et celui-ci, ne se distribuant pas aléatoirement, se concentre sur les populations aux caractéristiques sociales plus faibles (Pan Ké Shon, 2002).

Mais le principal reproche est d'instituer une seule norme quantitative indifférenciée des relations en omettant de la moduler par les qualités de celles-ci. Quelles sont donc ces qualités qui sont attribuées aux relations ? Au préalable, on peut affirmer que les qualités des relations sont pluridimensionnelles et on en dénombre au moins trois. D'abord, les relations, en terme de déprivation, sont souvent perçues à travers l'impossibilité de recours aux ressources matérielles des partenaires (Dickes, 1992). Ainsi, les aides diverses comme le prêt d'argent, le don, l'hébergement éventuel, et plus généralement la mise à disposition de biens et de services représentent effectivement des éléments influant sur le bien-être des personnes. Ensuite, les relations interpersonnelles sont source d'informations précieuses pour trouver un emploi (Degenne et alii, 1991, Forsé, 1994, 2000, Giret et alii, 1996, Marry, 1983), obtenir un conseil éclairé pour un investissement, orienter la carrière ou le cursus des enfants, etc. Il y a, *a priori*, des disparités d'accès à ces ressources accentuant ou compensant les autres inégalités

sociales. En fait, c'est le concept de capital social qui hante les esprits des chercheurs qui valorisent les disparités de ressources et uniquement celles-ci.

Au-delà des aspects matériels, la relation aux autres s'éprouve d'abord par l'intensité du lien, les relations de confiance (Ferrand, Mounier, 1993), l'appui psychologique dans un monde où « *l'individu est incertain* » et les dépressions courantes (Ehrenberg, 1998). Ainsi, l'isolement relationnel peut déboucher sur une perte de l'estime de soi et des pathologies psychologiquement douloureuses, d'où découlent une surconsommation d'alcool et de psychotropes. Les personnes isolées sont alors prises dans un engrenage car elles développent une incapacité à recréer du lien aboutissant sur l'enfermement et dans les cas extrêmes au suicide (Kaufmann, 1994). La part « purement » relationnelle participe à l'évidence de la construction de soi et entre dans la constitution de l'équilibre, du bien-être immatériel mais cependant bien réel des individus. Ainsi, l'isolement relationnel révèle un « certain » mal-être couplé fréquemment avec le sentiment de solitude, l'impression d'ennui et touchant des populations socialement fragiles (Pan Ké Shon, 2003). Cet aspect primordial des relations semble délaissé notamment par les auteurs se préoccupant des critères de pauvreté des conditions de vie.

Bien sûr, chaque type de relation possède une intensité relationnelle propre et un ami n'est pas équivalent avec un voisin (quand ils ne se confondent pas) et en ce sens une pondération devrait intervenir pour chaque type d'interlocuteur dans la construction de l'indicateur. Mais il se pose alors un problème quasi insoluble : comment définir ces pondérations et à partir de quels critères ? Un ami vaut combien de voisins et ce dernier combien de membre de la famille éloignée ou non ? De plus, chaque individu ayant des affinités personnelles propres, cette solution ajouterait un biais normatif plutôt qu'un véritable gain de précision. L'intensité relationnelle semble se dérober à la mesure. C'est pour cette raison et plus profondément pour la répulsion qu'inspire les éléments psychologisant aux économistes et même aux sociologues que les études relatives à la pauvreté relationnelle occultent cette dimension.

Pour revenir aux indicateurs de ressources matérielles ou d'informations provenant de tiers en cas de besoin, ils n'échappent pas à toutes critiques car ils font fréquemment référence au potentiel d'aide supposé qui ne correspond pas forcément aux aides qui seraient réellement obtenues (par exemple : « En cas de coup dur, pourriez-vous obtenir d'une de vos relations... »). L'indicateur se révèle alors taxable de subjectivité et entaché de soupçon de biais. Pour les aides effectives, en dehors des étudiants et des jeunes, on suppose qu'elles manifestent une corrélation étroite avec les indicateurs de pauvreté monétaire puisque ceux nécessitant une aide font aussi partie des plus démunis, ce qui reviendrait à mesurer grosso modo la même chose. En fait, le principal problème de l'utilisation d'un indicateur des aides effectives est que celui-ci indifférencie ceux qui n'en ont pas besoin et ceux qui malgré ce besoin ne trouvent pas d'aidant. Ainsi par exemple, les jeunes diplômés recherchant un emploi sollicitent moins leur réseau de relations car ils n'en ressentent pas la nécessité alors qu'une part des personnes ne pouvant pas valoriser un diplôme ne trouvent pas non plus d'appui à cause d'un réseau de relations déficient ou peu efficient. Par ailleurs, la non-concordance entre intensité du lien et pertinence des aides d'information est probable si l'on en croit l'hypothèse de la force des liens faibles avancée par Granovetter (1973) ou des trous structuraux développée par Burt (1992), ce qui indiquerait que les avantages de bien-être psychologique ne sont pas toujours dispensés par les mêmes partenaires que ceux procurant les aides les plus productives. Au travers de ces quelques remarques, il est perceptible que non seulement les qualités des liens sont multiples mais que ces qualités ne se concentrent pas sur les mêmes types de partenaires. Pour ces raisons et de façon paradoxale, l'indicateur global de contacts réels menés au cours d'une période, et faute de mieux, conserve sa validité.

Enfin, contrairement à la pauvreté monétaire, la pauvreté relationnelle est moins dépendante (du capital relationnel) du ménage que de l'individu. En effet, les contacts des parents ne sont pas ceux des enfants et plus généralement les relations d'un membre du ménage ne sont pas forcément partagées avec les autres membres (voir par exemple les différences du nombre de relations entre hommes et femmes en couples, Blanpain, Pan Ké Shon, 1999a). De ce fait, c'est moins le nombre d'individus

pauvres relationnellement qui prend sens (celui-ci dépend mécaniquement du seuil retenu) que le différentiel entre les populations évaluées.

3. Construction de l'indicateur d'isolement relationnel

Conventions

Comme pour la pauvreté monétaire, l'aspect conventionnel se porte aussi sur le seuil à retenir. Tout d'abord, il y a un choix à opérer entre isolement «absolu», c'est-à-dire les personnes n'ayant eu aucune relation pendant la période de référence (une semaine) ou isolement relatif. L'isolement relationnel absolu ne concerne qu'une très petite part de notre échantillon puisque seules 44 personnes sur un total de 11 802 répondants sont concernées. Ce qui exclu de fait l'utilisation de cet indicateur pour sa valeur instrumentale réduite et pour la construction d'un estimateur post-stratifié. Il est nécessaire alors de définir le seuil de l'indicateur d'isolement relatif.

Auparavant, il nous faut établir l'indicateur global de relations. Quatre questions sont disponibles dans l'enquête « Vie de quartier » :

- Au cours des 8 derniers jours, avec combien d'amis différents avez-vous discuté, en dehors de conversations téléphoniques ?
- Au cours des 8 derniers jours, avec combien de voisins différents avez-vous discuté, en dehors de conversations téléphoniques ? (en dehors de simples salutations et en dehors des personnes citées comme amis)
- Au cours des 8 derniers jours, avec combien de membres de votre famille hors ménage avez-vous discuté, en dehors de conversations téléphoniques ?
- Au cours des 8 derniers jours, combien de coups de téléphone personnels avez-vous pu donner ou recevoir environ ?

Pour cette dernière question, à chaque tranche est attribuée un équivalent contact visuel partant du fait que ces contacts peuvent être répétés avec le même individu et sont peut-être tendanciellement moins forts. Cette question n'est pas totalement satisfaisante puisqu'on mélange la fréquence des contacts avec le nombre des différents d'interlocuteurs. De plus, si l'utilisation de tranches s'avère pertinente pour la collecte de ces données, elles induisent une imprécision mais qui est tempérée par la modestie de la prise en compte. Malgré ces remarques et compte tenu de l'importance prise par la sociabilité téléphonique (Rivière, 2001) nous avons tenu à intégrer ces informations dans notre indicateur de la façon suivante :

Nombre de coups de téléphones reçus ou donnés :

- 0 → 0 contact
- De 1 à 5 → 1 contact
- De 6 à 10 → 2 contacts
- De 11 à 20 → 3 contacts
- Plus de 20 → 4 contacts

L'indicateur de totalisation des contacts s'obtient donc en sommant pour chaque personne le total de ses contacts hebdomadaires y compris téléphoniques comme précisé ci-dessus. Il est à noter que les questions s'efforcent à l'objectivité en recensant des contacts effectifs, réalisés au cours de la semaine précédent l'interrogation et ne fait pas appel à l'estimation du volume du réseau relationnel qui est plus sujet à la subjectivité, à une évaluation d'un réseau fantasmé plus que réel, notamment chez les personnes âgées (Héran, 1988).

Remarques

Par souci de limitation de la durée de passation du questionnaire et de son équilibre, le protocole d'enquête de Vie de quartier exclut les relations avec les collègues et les «autres relations», c'est-à-dire inconnus, relations des enfants, parents d'élèves, etc. Le choix de l'exclusion s'est porté sur les

collègues, à cause de la seule sélection des actifs occupés induite par la nature de la relation et de la dépendance du nombre potentiel de relations avec les collègues selon la taille de l'entreprise. La non prise en compte des contacts professionnels influe surtout sur le nombre de relations des actifs de 25 à 55 ans et réduit conséquemment le seuil retenu de l'isolement relationnel.

Pour finir, l'indicateur privilégie les relations interpersonnelles informelles afin d'éviter d'intégrer, notamment les participations aux associations, car on mélangerait alors ce qui est du domaine de l'activité (par exemple : je participe parce que je veux faire du sport, parce que je m'intéresse aux études des enfants) avec la sociabilité volontaire ou informelle (j'ai une conversation avec un membre, non comme but de mon adhésion mais comme bénéfice marginal), même si la ligne de partage peut s'avérer mobile et plus floue.

Dispersion

La variable du total des relations créée, il reste à observer la dispersion de ceux-ci avant d'aller plus loin.

Mean 14.12043
 Median 12.00000
 Mode 7.00000

quintiles

Isol01	isol05	isol10	isol15	isol20	isol25	isol30	isol35	isol40	isol45	isol50	isol55	isol60
1	3	4	5	6	7	8	9	10	11	12	13	14
isol65	isol70	isol75	isol80	isol85	isol90	isol95	isol99	isol100				
15	17	18	20	23	26	34	48	66				

Lecture : les deux derniers chiffres de la variable isolxx correspondent aux déciles et le chiffre juste en dessous correspond au nombre maximum de contacts hebdomadaires. Exemple : isol10 correspond au 1^{er} décile soit 4 contacts au plus par semaine.

Effectifs des isolés selon les différents seuils

Les deux derniers chiffres correspondent aux nombres de contacts. Exemple : isolement04 désignent les personnes ayant eu 4 contacts ou moins au cours de la semaine de référence.

isolement00	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11758	99.63	11758	99.63
1	44	0.37	11802	100.00

isolement01	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11615	98.42	11615	98.42
1	187	1.58	11802	100.00

isolement02	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11336	96.05	11336	96.05
1	466	3.95	11802	100.00

isolement03	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	10960	92.87	10960	92.87
1	842	7.13	11802	100.00

isolement04	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	10491	88.89	10491	88.89
1	1311	11.11	11802	100.00

isolement05			Cumulative	Cumulative
	Frequency	Percent	Frequency	Percent
0	9934	84.17	9934	84.17
1	1868	15.83	11802	100.00

isolement06			Cumulative	Cumulative
	Frequency	Percent	Frequency	Percent
0	9279	78.62	9279	78.62
1	2523	21.38	11802	100.00

isolement07			Cumulative	Cumulative
	Frequency	Percent	Frequency	Percent
0	8578	72.68	8578	72.68
1	3224	27.32	11802	100.00

Frequency Missing = 117

Au lieu de calquer notre indicateur d'isolement exactement sur celui de la pauvreté monétaire ou de chercher la « vraie référence », ce qui est autant d'une rigueur et d'une objectivité illusoire que tout autre choix, nous allons choisir un indicateur répondant à trois qualités : il est nécessaire qu'il possède un effectif suffisant pour les traitements à venir et de ce fait évite autant que possible les aléas d'enquête mais aussi qu'il se montre le plus précis et stable possible lors de l'analyse logistique. La précision sera évaluée en se référant au nombre de paires concordantes. Quant à la stabilité, c'est celle relative aux autres seuils d'isolement : les composantes du modèle de l'indicateur retenu doivent tendanciellement rester stables quelque soit le seuil choisi. Après différentes itérations, le seuil montrant ces qualités est celui du premier décile, soit avoir eu un contact avec au plus 4 individus différents au cours de la semaine de référence.

Les zones urbaines sensibles

La classification de quartiers en zones urbaines sensibles obéit d'abord à des règles administratives. La liste, actuellement 751 Zus, est arrêtée par la Délégation Interministérielle à la Ville (DIV) en accord avec les préfetures et les collectivités locales. Elle privilégie les grands quartiers périphériques, en fonction de leurs profils socio-économiques évalués par l'écart de leur situation à la ville ou à l'agglomération d'appartenance et des priorités dans les « contrats de ville ». Ce n'est donc pas une indication permettant d'établir une échelle de précarité nationale des quartiers. En ce sens, une zone urbaine sensible d'une région pourrait très bien être déclassée si elle avait appartenu à une autre rencontrant plus de problèmes sociaux.

4. Présentation des sources

L'enquête « Vie de quartier » a été réalisée en partenariat avec la Délégation Interministérielle à la Ville (DIV), avec le Plan Urbanisme Conception Architecture (PUCA) du ministère de l'Équipement, avec l'Observatoire national de la pauvreté et de l'exclusion sociale, avec la Dares et la Drees du ministère du travail et de la santé, avec l'Union nationale des HLM, l'Institut des Hautes Etudes sur la Sécurité Intérieure (IHESI) du ministère de l'Intérieur et la Caisse Nationale d'Allocations Familiales (CNAF). La collecte s'est déroulée d'avril à juin 2001 auprès de 10 033 ménages et 11 500 individus répondants. Cette enquête s'insère dans le cadre du dispositif d'Enquêtes Permanentes sur les Conditions de Vie des ménages (EPCV) établi par l'Insee et a pour objet l'étude des disparités dans les divers types de quartiers. La représentativité uniquement nationale de l'échantillon ne permet donc pas d'obtenir des données fiables au niveau infra national.

Les répondants sont âgés de 15 ans et plus, l'étude ne porte donc que sur cette population.

Surface couverte des Zus par les 3 extensions VdQ de Bretagne

(Quartiers prioritaires de la politique de la ville) %

	Surface couverte de la Zus
Maurepas	96
Le Blosne	57
La Découverte	34
Ensemble	61,5

Source : DR Insee Bretagne

Lecture : la partie en Zus du quartier représente 96 % des habitants.

Par ailleurs, une extension de l'enquête Vie de quartier a été réalisée à l'identique par la DR de Bretagne sur deux quartiers prioritaires de la politique de la ville de Rennes (Le Blosne et Maurepas) et un quartier de Saint-Malo (La Découverte), ces quartiers incluant chacun une Zus (Voir cartes en annexes 1 et 2). Les données de Bretagne nous permettront de confronter notre estimateur petit domaine aux estimations directes issues de la collecte régionale. Les données concernant la composition sociodémographique du petit domaine (les 3 Zus de Bretagne) proviennent du RP99.

5. Principes généraux

On part de l'hypothèse (à vérifier ensuite) que l'isolement relationnel est identique sur l'ensemble du territoire métropolitain, seule la composition sociale des divers niveaux socio-spatiaux change et explique les disparités de niveau du nombre d'isolés. En s'appuyant, d'une part sur la moyenne nationale de chaque caractéristique « expliquant » l'isolement et d'autre part sur la composition exacte de ces niveaux socio-spatiaux (un groupe de Zus par exemple) selon les caractéristiques précédemment citées, il est possible de déduire la moyenne du nombre d'isolés de ce découpage socio-spatial.

Les grands traits de la marche à suivre sont alors simples à tracer :

1. Établir un modèle de comportement individuel à partir de l'enquête nationale « Vie de Quartier » ;
2. Vérifier que le comportement individuel national est identique au comportement des personnes en ZUS, « toutes choses étant égales par ailleurs » ;
3. Déterminer les variables et modalités les plus actives ;
4. Celluler à l'identique, selon les informations précédentes, les données disponibles VdQ et RP99 à partir des variables de comportement ;
5. Calculer la moyenne de chaque cellule VdQ ;
6. Calculer les coefficients de structure de chaque cellule de la population du domaine à étudier, obtenus grâce au RP99 ;
7. Repondérer les moyennes des cellules VdQ par les coefficients de structure ;
8. Sommer ;
9. Contrôler les résultats.

L'estimateur petit domaine ainsi produit s'appelle **estimateur post-stratifié** à cause de la repondération des effectifs de chaque cellule effectuée *a posteriori*. D'autres méthodes d'estimation pour petits domaines sont applicables suivant la variable à estimer et les données disponibles, pour une revue générale de ces techniques voir Destandeu, 1996 ; Singh et alii, 1994, Hidioglou, 1992.

6. Formalisation de la méthode employée

L'estimation directe de l'isolement relationnel en France métropolitaine s'appuie sur l'estimateur d'Horvitz-Thompson, ou plus simplement la moyenne pondérée :

$$\hat{y}_h = \frac{\sum_{k \in h_{VdQ}} w_k Y_k}{\sum_{k \in h_{VdQ}} w_k}$$

où w_k représente le poids de sondage de l'individu k .

C'est cet estimateur qui nous sert de référence pour l'isolement sur 3 Zus de Rennes (2) et Saint-Malo (1) grâce aux extensions régionales de Bretagne de l'enquête Vie de quartier.

L'expression de l'estimateur « petit domaine ou post-stratifié »

L'estimation de l'isolement Y dans le petit domaine z est égale à la somme, pour chaque cellule h , de l'isolement nationale moyen dans la cellule h , pondérée par l'effectif de cette cellule dans le petit domaine z . Ainsi, l'expression de l'estimateur « petit domaine » de Y dans le domaine d'intérêt z (les 3 Zus de Bretagne) s'écrit :

$$\hat{y}_z = \sum_{h=1}^x \frac{N_{h,z}}{N_z} \bar{y}_h$$

La moyenne \bar{Y}_h est estimée à partir de l'échantillon VdQ national.

x représente le nombre total de cellules, variant selon les différentes versions de l'estimateur.

$\frac{N_{h,z}}{N_z}$ est le coefficient de structure de chaque cellule dans les 3 Zus de Bretagne z , tiré de l'enquête VdQ.

7. Le modèle initial de comportement individuel

D'après les études disponibles, l'isolement relationnel répondrait à trois caractéristiques principales qui ne sont pas totalement sans influences les unes sur les autres (voir la bibliographie sélective en fin de document), auxquelles s'ajouteraient des éléments extérieurs comme l'influence possible de l'environnement ou du contexte, l'effet propre du quartier, le comportement original régional, etc. Le modèle de comportement individuel répondrait aux points suivants qui ne sont pas sans interrelations entre eux :

Le cycle de vie perçu à partir de l'âge et de la situation familiale. Au fur et à mesure de l'avancée en âge, les individus ont de moins en moins de relations, moins de disponibilités, recentrage sur la cellule familiale à l'âge mûr, disparition des relations professionnelles à la retraite, décès et handicaps pénalisants expliquent la raréfaction des contacts (Blanpain, Pan Ké Shon, 1999b).

La position sociale qui demeure un facteur explicatif au travers des revenus, de la PCS, habiter en cité, voire le pays de naissance qui traduit parfois plus la hiérarchie sociale qu'un comportement culturel original.

Le capital culturel ou plus exactement le **capital scolaire** puisqu'il est approché grâce au diplôme et qu'ils ne se confondent pas totalement. Il est possible d'inclure les participations à une association, religieuse, politique, syndicale, culturelle, sportive qui sont autant de comportements voire d'atouts culturels. De plus, l'activité associative a pour bénéfice, plus ou moins indirecte, de créer de nouveaux et de plus nombreux contacts. Par ailleurs, on peut considérer que le capital scolaire n'est qu'une composante de la position sociale à cause de

la corrélation étroite entre le diplôme et la hiérarchie sociale. Néanmoins, nous avons préféré les distinguer à cause de l'influence particulière occupée par le diplôme sur la sociabilité des individus (Héran, 1988) et pour la suite des traitements.

La contrainte physique exercée par les handicaps invalidants.

A ces quatre points, il faut rajouter un terme d'hétérogénéité individuelle et de caractéristiques inobservées.

Tableau 1 - Probabilité d'être isolé relationnellement

		Paramètre	Seuil
Constante		-2.1156	<.0001
Âge	15-19 ans	-0.9335	<.0001
	20-24 ans	-0.6087	<.0001
	25-34 ans	-0.3870	0.0002
	35-69 ans	Réf.	-
	70 ans et plus	0.4908	<.0001
Diplôme	Aucun	0.3793	<.0001
	Certificat, BEPC	0.1834	0.0271
	CAP/BEP, Bac, Supérieur (1)	Réf.	-
	Bac +2	-0.1753	0.1548
Activité	Agric., artisan, comm. cadre, prof. inter., employé, OS, chômeur (1)	Réf.	-
	Chef d'entreprise, étudiant, OHQ	-0.3176	0.0027
	Autres CS, retraité	0.1942	0.0333
Type de ménage	Seul et veuf	-0.1416	0.2199
	Couple sans enfant	-0.2150	0.0095
	Enfant du couple	-0.3146	0.1214
	Famille monoparentale	0.2108	0.0532
	Autre ménage (1)	Réf.	-
Pays de naissance	Pays d'Afrique	0.4998	<.0001
	France et autres pays (1)	Réf.	-
Nombre de personnes du ménage	Moins de six	Réf.	-
	6 et plus	0.4412	0.0047
Religion	Ni pratique, ni sentiment religieux	0.2864	<.0001
	Refus de répondre	0.1828	0.2255
	Pratique régulière	-0.2117	0.0304
	Pratique occasionnelle ou sentiment d'appartenance	Réf.	-
Associations	Civique	-0.3462	0.0079
	Anciens combattants	-0.4822	0.0020
	Sportive	-0.5240	<.0001
	Culturelle	-0.5835	<.0001
	Non appartenance	Réf.	-
Type de triris - axe1 (2)	Décile 1 à 2	-0.2282	0.0050
	Décile 3 à 4	-0.2665	0.0010
	Autres déciles	Réf.	-
Revenus par unité de consommation	00 à 20 KF	0.5338	0.0109
	20 à 30 KF	0.3743	0.0152
	30 à 50 KF	0.3311	<.0001
	50 à 240 KF	Réf.	-
	Supérieurs à 240 KF	0.2935	0.0664
Habitat en cité	Oui	0.1348	0.1429
	Non	Réf.	-
Handicap permanent empêchant de travailler	Oui	0.4074	0.0068
	Non	Réf.	-
Quartier en Zus	Oui	0.2422	0.0206
	Non	Réf.	-

(1) Les modalités ne montrant pas de différence avec la référence ont été regroupées avec celle-ci.

(2) Un triris est une division spatiale du territoire regroupant 3 iris dont chacun représente environ 2000 personnes. La typologie socio-économique des quartiers 1999 regroupe les quartiers ou les communes en fonction du profil socio-économique de leurs habitants (emploi et secteur d'activité). Elle fait ressortir 3 oppositions : la richesse à la pauvreté ; le chômage à l'activité ; les activités tertiaires et l'artisanat à l'industrie et le technique. Ce sont les valeurs de la première opposition qui ont servi ici dans la définition des triris. Les premiers déciles affectent les quartiers les plus aisés.

Le logit indique près de 69 % de paires concordantes ce qui sans être spectaculaire paraît acceptable (modèle complet en annexe 3). Le logit montre aussi que plusieurs variables sortant avec force sont indisponibles dans les fichiers du RP : ainsi les associations, le sentiment religieux, les revenus, et dans une moindre mesure le handicap (tableau 1 et annexe 3). L'abandon de ces variables lors du cellulaire à partir du RP laisse présager une perte dommageable pour la suite du traitement.

Cependant, par le jeu des corrélations entre variables, il est possible d'observer qu'elles sont les variables qui captent tout ou partie de la valeur abandonnée par le retrait des variables absentes dans les fichiers du RP99 (tableau 2 et annexes 3 et 4). Le tableau rend aisé la lecture de la captation par la variable de diplômes (colonne B) de la force explicative abandonnée par les variables d'associations, le solde ne trouvant pas de variable attractive vient alors gonfler la référence. L'absence des variables d'associations ne devrait donc pas être trop pénalisant dans le modèle contracté pour la détermination des cellules (cf. infra).

Tableau 2 - Comparaisons de la force explicative de chaque variable entre modèles complet et contracté

	Wald Chi2		
	Modèle complet A	Modèle RP contracté B	Différence B - A
Associations	63.1540	-	-
Age	63.0623	48.8406	-14,2217
Sentiment religieux	29.0237	-	-
Diplômes	25.7316	60.0882	34,3566
Revenus par U.C.	24.7495	-	-
Pays de naissance	20.4499	20.1863	-0,2636
Pcs	16.3780	9.5091	-6,8689
Type de ménage	14.8466	25.0825	10,2359
Axe1 typo. Tabard-Houssard	14.5905	10.8058	-3,7847
Famille nombreuse (>=6)	7.9942	11.9277	3,9335
Handicap empêchant le travail	7.3217	-	-
Type d'habitat	2.1469	3.5029	1,356
Référence (intercept)	682	1032	-

Reste à déterminer, si le modèle initial s'applique aux personnes habitant dans une Zus ou si la concentration de personnes aux caractéristiques sociales plus faibles dans un même lieu, souvent vécu comme stigmatisant, crée un effet dynamique propre au quartier et l'en distingue des autres quartiers. Dit autrement, il est nécessaire de vérifier qu'habiter en zone urbaine sensible ne conduit pas à l'enfermement sur le quartier, dans le logement et à éviter tout contact à cause notamment de l'insécurité plus grande dans les quartiers défavorisés. D'ores et déjà, le modèle initial indique un effet significatif du fait d'habiter en Zus sur l'isolement. Cependant, on ne peut se contenter de ce résultat car il est possible que cette variable capte des caractéristiques individuelles particulièrement faibles des personnes vivant en Zus que les autres variables, revenus, pcs, diplômes, etc. peinent à percevoir. Par exemple, on peut être amené à penser que les personnes ayant le baccalauréat et vivant en Zus possèdent des caractéristiques individuelles plus faibles que celles ayant le bac mais habitant hors Zus, ne serait-ce déjà par les inégalités entre les divers bacs.

8. Vérification d'un éventuel effet quartier en Zus - Surmodèle isolement relationnel

L'effet quartier

La question est souvent posée de savoir si seules les caractéristiques sociales des personnes « expliquent » le comportement individuel ou si habiter un quartier défavorisé aggrave les disparités,

en l'occurrence l'isolement (pour une revue de l'effet quartier voir Marpsat, 1999). Il est possible de tester la significativité de la différence des valeurs des paramètres du modèle national hors Zus et du modèle Zus de l'enquête Vie de quartier afin de vérifier si les différences entre variables identiques de ces deux modèles sont robustes. L'idée est ici de pouvoir comparer le comportement des personnes ne vivant pas en Zus de celles vivant en Zus en évaluant un éventuel « effet spécifiquement Zus ». Si on ne peut mettre à jour un effet local, il n'y aurait pas alors de correctif dans l'estimation de l'isolement dans les Zus.

Surmodèle isolement relationnel

Pour cela, au lieu d'établir deux équations de régression logistique Zus & non Zus, on établit une seule équation en dichotomisant chaque variable pour chaque zone d'intérêt (ex. : « femme et zfemme », où Femme représente la variable dichotomique femme ne vivant pas en Zus et zfemme être une femme vivant en Zus) et une indicatrice pour les Zus est introduite (Godechot, 2000). Nous lui empruntons la méthode et le texte suivant :

« Formellement, on peut écrire les deux régressions logistiques aux deux dates comme deux modèles pour deux populations :

$$P(y_{it_1} = 1) = [1 + \exp[-(a_{t_1} + \sum_{j=1}^{j=n} b_{jt_1} x_{jit_1} + u_{it_1})]]^{-1}$$

et

$$P(y_{it_2} = 1) = [1 + \exp[-(a_{t_2} + \sum_{j=1}^{j=n} b_{jt_2} x_{jit_2} + u_{it_2})]]^{-1}$$

où y_{it_1} représente l'événement étudié, a_{t_1} la constante du modèle à la date t_1 , b_{jt_1} le paramètre estimé associé à la variable x_j à la date t_1 , x_{jit_1} la valeur de la variable x_j pour l'individu i à la date t_1 , u_{it_1} le résidu du modèle pour l'individu i à la date t_1 .

Pour faciliter la comparaison des paramètres et établir l'éventuelle significativité de la différence ($b_{jt_1} - b_{jt_2}$), il est plus commode d'estimer un "surmodèle" plutôt que d'estimer séparément les deux équations. Pour cela on "empile" les deux enquêtes et on estime :

$$P(y_i = 1) = [1 + \exp[-(a + a' \cdot 1_{t_2} + \sum_{j=1}^{j=n} b_{jt_1} x_{jit_1} + \sum_{j=1}^{j=n} b_{jt_2} x_{jit_2} + u_i)]]^{-1}$$

où 1_{t_2} signifie une indicatrice pour la période 2. (La valeur de la constante pour la date 1 est égale à a et pour la période 2, elle est égale à $a + a'$). Cette dernière estimation est équivalente à la première, mais elle permet de faire des tests rigoureux de différence des paramètres aux deux dates. »

Ici, il n'est pas utile d'empiler les enquêtes, il suffit de considérer que nous sommes en présence de 2 échantillons distincts. On peut ainsi tester aisément la significativité de la différence des paramètres entre le modèle national hors Zus et le même modèle mais en Zus. Le résultat figure dans le tableau 3. La lecture de ce tableau montre d'abord que les variables du modèle Zus sont moins fréquemment et moins fortement significatives que celles du modèle hors Zus, la faiblesse de l'échantillon Zus en est bien sûr la cause. Ces différences de niveau sont-elles réellement significatives ? Voilà la question à laquelle il nous faut répondre. La dernière colonne du tableau 3 exprime le seuil de significativité de cette différence (voir annexe 5 pour la syntaxe SAS).

Seules les modalités possède un CEP ou un BEPC et les personnes non croyantes manifestent un écart significativement différent au seuil de 5 % dans ces deux types de quartiers. La très grande majorité des autres coefficients traduisent une absence nette de significativité des différences. Ce manque de significativité des différences exprime l'absence d'effet spécifique quartier en Zus sur l'isolement relationnel, au moins à partir des données de l'enquête Vie de quartier. Pour autant, la variable Zus exprime une différence manifeste sur l'isolement avec les habitants hors Zus comme le montre le tableau 3b et qui s'interprète alors comme l'effet de la concentration d'individus aux caractéristiques sociales plus faibles corrélées à l'isolement. L'effet perceptible de la variable Zus dans le logit initial

(tableau 1) est donc à mettre au compte de caractéristiques individuelles plus faibles et non mises à jour que les variables sociodémographiques traditionnelles ne parviennent pas à capter.

Tableau 3 - Significativité des différences suivant l'habitat en Zus ou hors Zus

	Etre isolé relationnellement	Hors Zus		En Zus (1)		Significativité Pr > Chi2 ? B _i
		Coeff.	Pr > Chi2	Coeff.	Pr > Chi2	
Constante		-2.3611	<.0001	-2.3611	<.0001	-
Zone d'estimation	En Zus	-	-	0.0345	0.9170	-
	Hors Zus	-	-	-	-	-
Age	15-19 ans	-1.0508	<.0001	-0.7912	0.0750	0.6080
	20-24 ans	-0.7000	<.0001	-0.1450	0.6490	0.1179
	25-34 ans	-0.3866	0.0007	-0.2745	0.2509	0.6726
	35-69 ans	Réf.	-	Réf.	-	-
	70 ans et plus	0.5051	<.0001	0.1698	0.6438	0.3800
Diplôme	Aucun	0.3873	<.0001	0.6114	0.0062	0.3535
	Certificat, BEPC	0.1179	0.1864	0.7658	0.0013	0.0110
	Bac +2	-0.1801	0.1603	-0.1228	0.7942	0.9066
	CAP/BEP, Bac, Supérieur (1)	Réf.	-	Réf.	-	-
Activité	Agric.. artisan, commerçant, retraité, inactifs	0.4572	<.0001	0.1580	0.5230	0.2689
	Cadre. profession intermédiaire, employé,	0.2822	0.0038	0.0566	0.7856	0.3263
	Chômeur. étudiant, ouvrier, Chef d'entreprise (1)	Réf.	-	Réf.	-	-
Type de ménage	Seul et veuf	-0.1012	0.4032	-0.5623	0.1700	0.2805
	Couple sans enfant	-0.2214	0.0113	-0.1001	0.6964	0.6547
	Enfant du couple	-0.1988	0.3824	-0.8938	0.0560	0.1815
	Famille monoparentale	0.2298	0.0611	0.1944	0.4251	0.8968
	Autre ménage (1)	Réf.	-	-	-	-
Pays de naissance	Pays d'Afrique	0.5172	<.0001	0.5996	0.0059	0.7455
	France et autres pays (1)	Réf.	-	Réf.	-	-
Nbre de pers. du ménage	Moins de six	Réf.	-	Réf.	-	-
	6 et plus	0.4794	0.0094	0.5454	0.0761	0.8538
Religion	Ni pratique, ni sentiment religieux	0.2326	0.0017	0.7230	0.0002	0.0190
	Refus de répondre	0.1269	0.4383	0.6046	0.1343	0.2728
	Pratique régulière	-0.2122	0.0461	-0.1826	0.4721	0.9144
	Pratique occasionnelle ou sentiment d'appartenance	Réf.	-	Réf.	-	-
Associations	Civique	-0.2989	0.0263	-1.0570	0.0511	0.1745
	Anciens combattants	-0.4724	0.0030	-0.6040	0.4389	0.8688
	Sportive	-0.5412	<.0001	-0.3888	0.3689	0.7347
	Culturelle	-0.5714	<.0001	-0.8708	0.0808	0.5635
	Non appartenance	Réf.	-	Réf.	-	-
Type de triris axe1 (2)	Décile 1 à 2	-0.2215	0.0136	-0.3418	0.0909	0.5863
	Décile 3 à 4	-0.2090	0.0181	-0.5971	0.0043	0.0871
	Autres déciles	Réf.	-	Réf.	-	-
Revenus par unité de consommation	00 à 20 KF	0.6174	0.0078	0.1811	0.7260	0.4414
	20 à 30 KF	0.4992	0.0045	0.1004	0.7729	0.3061
	30 à 50 KF	0.3502	0.0002	0.2229	0.2830	0.5755
	50 à 240 KF	Réf.	-	Réf.	-	-
	Supérieurs à 240 KF	0.2608	0.1072	0.8584	0.4810	0.6267
Habitat en cité	Oui	0.0989	0.3709	0.3344	0.0637	0.2655
	Non	Réf.	-	Réf.	-	-
Handicap permanent	Oui	0.3716	0.0296	0.8685	0.0118	0.1968
	Non	Réf.	-	Réf.	-	-

Lecture : Dans l'équation de régression, les jeunes vivant hors Zus de 15 à 19 ans ont une probabilité significative (coeff. - 1.0508) de ne pas être isolés relationnellement par rapport aux personnes de 35 ans à 69 ans. Quant aux jeunes de 15 à 19 ans vivant en Zus, ils ont une probabilité un peu moindre de ne pas être isolés relationnellement (-0.7912). Ces deux paramètres ne sont pas significativement différents entre eux (dernière colonne).

La significativité de la différences des paramètres se lit de la façon suivante: la différence des coefficients entre le modèle en ZUS et le modèle hors ZUS est significatif au seuil de 1 % pour les personnes ayant un certificat d'études primaires ou un BEPC et pour les personnes n'ayant ni pratique ni sentiment religieux. A 8 % pour les personnes habitant un quartier faisant partie d'un quartier du 3^e au 4^e décile ; on considère comme non significatif les valeurs de cette dernière colonne prises au-delà de 0,1000 qui correspond au seuil de 10 %.

** La valeur de la constante du modèle Zus s'obtient en additionnant la constante plus la valeur du paramètre « en Zus ».

Tableau 3b - Isolement suivant l'habitat en Zus ou non

	Non Zus	En Zus
Non isolés relationnellement	89.6	84.1
Isolés relationnellement	10.4	15.9
Total	100	100

Source : VdQ

9. Hiérarchie des variables dans l'isolement relationnel

La menace de biais venant d'un effet quartier en Zus étant écartée, nous pouvons passer à l'étape suivante, c'est-à-dire la détermination des variables et des modalités à retenir pour effectuer le cellulaire.

Chaque cellule servira donc à calculer :

1. La moyenne de l'isolement dans la cellule définie à partir de VdQ ;
2. Les effectifs de cette cellule dans le petit domaine nous intéressant, soit les 3 Zus de Bretagne.

Un nombre de cellules important permettra que chacune soit finement définie mais plus le nombre sera important et moins les effectifs seront nombreux dans chacune d'elles entraînant une plus grande imprécision. Dans le cas qui nous concerne, le fichier Vie de quartier contient 11 802 individus ayant répondu aux questions permettant de définir l'indicateur d'isolement. Si on vise un seuil moyen de 100 personnes par cellule, nous obtenons un nombre de cellules maximum de 118. Sachant que les effectifs seront inégaux dans chaque cellule, nous aurons à opérer un regroupement des cellules dégarnies « manuellement » en observant chaque cellule.

En exemple et pour fixer les idées, nous pourrions avoir 4 ou 6 variables dont les modalités se combinaient ainsi :

$$3 \times 3 \times 3 \times 4 = 108$$

$$2 \times 2 \times 2 \times 2 \times 3 = 96 \text{ etc.}$$

Pour ce dernier cas, l'introduction d'une nouvelle variable à deux modalités multiplie par deux le nombre de cellules soit 192 cellules ! On arrive donc vite aux limites possibles autorisées par l'échantillon de l'enquête. Ces exemples signalent que nous aurons à opérer un partage rigoureux entre le nombre de variables et le nombre des modalités servant au découpage des cellules. Comment effectuer ce choix cornélien ?

Revenons un peu en arrière. D'abord, la lecture du modèle contracté, c'est-à-dire du modèle réalisé à partir des variables communes à VdQ et RP99 (annexe 4) indique l'importance de la significativité de chaque modalité des variables mais pas l'importance résumée pour chaque variable du modèle (sexe, âge, diplôme, etc.). On souhaite connaître l'importance relative à accorder à ces diverses variables sur l'isolement relationnel afin de sélectionner les variables les plus pertinentes pour constituer les cellules servant à calculer l'estimateur du domaine et d'écarter les autres.

Hiérarchie des variables

En réalisant un test de nullité des variables, autrement dit en retirant une variable du modèle l'une après l'autre, on met à jour les différences relatives en mesurant la perte du χ^2 du modèle.

Cette opération s'effectue aisément en SAS en ajoutant après la liste des variables explicatives du logit l'instruction suivante :

```
age: test a1519=0, a2025=0, a2534=0, a7099=0 ;
```

```
diplomes:test diplo0=0, diplo1=0, diplo4=0 ;
```

```
run;
```

Où a1519 et diplo0 sont les noms des modalités des variables dichotomisées et âge et diplômes les labels optionnels des tests. Les modalités manquantes sont les modalités d'identification retirées du modèle logit. Le programme correspondant figure en annexe 4. Le tableau 4 indique de façon décroissante cette hiérarchie.

Il est à noter que l'introduction d'une autre variable significative dans le modèle fait fluctuer le chi2 des variables qui avaient capté auparavant tout ou partie de son pouvoir explicatif. Il ne faut donc pas s'attacher à une différence de quelques unités près. On peut considérer ainsi que « Famille nombreuse, axe 1 et Pcs » ont un pouvoir explicatif similaire.

Tableau 4 - Hiérarchie de la significativité des variables

Variables du modèle contracté	Wald chi2
Diplômes	60.0882
Age	48.8406
Type de ménage	25.0825
Pays de naissance	20.1863
Famille nombreuse (≥ 6)	11.9277
Axe1 typo. Tabard-Houssard	10.8058
Pcs	9.5091
Zus	6.2516

Lecture : Les diplômes montrent un pouvoir explicatif le plus important par rapport aux autres variables du modèle (en fonction de la perte du chi2).

Sélection des modalités

Les variables étant sélectionnées, leur nombre maximum déterminé, il faut maintenant les découper selon un nombre de modalités compatibles avec la taille de l'échantillon disponible, ici VdQ. L'objectif est d'obtenir un nombre de cellules avec chacune un effectif suffisant, permettant une estimation dotée d'une variance limitée. La difficulté principale à cette étape est de trouver un équilibre entre le nombre de variables utilisées et le nombre de modalités. Ainsi, l'utilisation d'une variable supplémentaire de croisement, si elle apporte de nouvelles informations, elle en détruit simultanément par la réduction des modalités induite. Et la seule sélection de 4 ou 5 variables délaie une part de la force explicative contenues par les autres variables corrélées à l'isolement : Famille nombreuse, axe 1, Pcs, etc. En final, les strates n'auront intégré qu'une part de la pertinence dégagée par le modèle.

Si le choix des variables accompli grâce à la significativité du chi2 est une méthode satisfaisante et aisée de mise en œuvre, par contre, la détermination des modalités ne bénéficie pas de la même rigueur, ni de la même aisance. Pour déterminer les bornes des modalités, la ligne de conduite suivie a été de sélectionner les modalités selon l'importance du coefficient de chaque modalité et également du Wald chi2 de la régression et de comparer les différentes possibilités de découpage de façon raisonnée. Les modalités aux coefficients proches sont agrégées. Ainsi, le nombre et le découpage des tranches a aussi été effectuée sous contrainte de laisser subsister suffisamment d'individus dans chaque case définie. De nombreux essais ont été réalisés afin d'obtenir la partition optimum.

Comme nous allons le voir, les estimations se révèlent très sensibles à la seule détermination fine des modalités puisqu'à partir des mêmes variables, la variation de l'estimateur petit domaine peut s'élever de plusieurs points. Ce qui montre l'importance particulière de cette phase et qui ne peut être négligée. La détermination des modalités reste néanmoins, l'un des points faibles de la démarche.

10. Constitution des cellules

Pratiquement, le cellulaire revient à créer une nouvelle variable fabriquée à partir des variables et des modalités sélectionnées. Cette variable « découpe » exhaustivement le fichier Vie de quartier en fonction des caractéristiques sociodémographiques mises à jour en autant de cellules. Elle sert également à découper à l'identique le fichier du RP du domaine étudié pour obtenir la structure sociodémographique de celui-ci. Chaque modalité de variable est dichotomisée, ce qui permet ensuite une fabrication et des rectifications plus aisées de la variable de cellulaire (annexe 7). Ensuite, les cellules dégarnies sont regroupées de façon logique et constituées afin qu'elles représentent un effectif suffisant. En effet, il est nécessaire de conserver suffisamment d'individus dans chaque case afin d'estimer une moyenne pas trop erratique et de limiter la variance de l'estimateur. Les variables et les modalités étant déterminées, il reste à effectuer le cellulaire en observant le comportement des différents découpages selon les variables utilisées (tableaux 5).

Le logit nous indique les variables influentes et la plus ou moins grande proximité des modalités introduites dans le modèle et jusqu'à présent nous sommes placés dans une logique de séparation des effets. Afin de déterminer un cellulaire le plus pertinent possible, nous allons essayer de nous placer dans une logique complémentaire, celle qui conduit à chaque étape du cellulaire d'utiliser les variables les plus actives.

Segmentation

La segmentation pourrait être l'outil approprié. Pour la tester nous avons fait appel à une macro adaptée en SAS par Stéfan Lollivier et qui précise :

« La macro SEGMENT réalise une segmentation de une ou plusieurs variables sur un certain nombre de critères explicatifs en utilisant un principe de décomposition au vu d'une stratification en deux sous-populations rendant maximale la part de variance expliquée. »

Si i ($= 1$ à n) désigne une modalité quelconque d'une caractéristique qui en comprend n et si n_i est le nombre de personnes possédant la modalité i et si x_i est le taux d'isolés parmi ces n_i la variance expliquée est définie par :

$$\sum n_i (x_i - x)^2$$

x étant le pourcentage d'isolés dans la population totale (D'après Dumontier, Guillaumat, 1996 ; On peut aussi consulter Saporta, 2000 ou Gelein, 2002).

Les estimateurs issus de la segmentation sont encore insatisfaisants car trop éloignés de l'estimation directe. Il est probable que la séparation binaire à chaque nœud de l'arborescence soit trop grossière, néanmoins elle suggère d'utiliser asymétriquement les variables afin de réaliser le découpage.

Les premières versions (1 à 4) s'appliquaient à utiliser de façon « uniforme » les différentes variables (tableaux 5 et annexe 6). A partir de la version numérotée 10, nous intégrons une variable composite : Africain ou ménage de ≥ 6 personnes justifiée par la proximité des coefficients du logit et par des effectifs réduits de ces 2 éléments. Enfin, pour la version 16, la variable appartenance en Zus est appliquée sur les cellules possédant un effectif encore suffisant. L'idée ici est de mobiliser le maximum des informations disponibles en fonction de la contrainte due aux effectifs restreints.

Tableaux 5 - Choix des variables et de leurs modalités

Version 1

Variables	Nbre de modalités	Modalités
Diplômes	4	Sans CEP, BEPC CAP, BAC, >bac+2 Bac +2
Age	4	15 à 19 ans 20 à 34 ans 35 à 69 ans 70 et plus
Type de ménage	3	Veuf, couple sans enfant, enfant d'un couple Monoparent adulte Autres cas
Pays de naissance	2	Pays d'Afrique Autres cas

Version 16

Variables	Nbre de modalités	Modalités
Diplômes	4	Sans CEP, BEPC CAP, BAC, >bac+2 Bac +2
Age	4	15 à 19 ans 20 à 34 ans 35 à 69 ans 70 et plus
Type de ménage	3	Célibataire, adulte d'un couple avec enfant Monoparental Veuf, séparé, couple sans enfant, enfant d'un couple, autre ménage
Pays de naissance ou ménage >= 6 personnes	2	Oui : Africain ou ménage de >=6 Non
PCS regroupée	3	Agric, artcom, AutrCS, retraité Cadres, profIn, emplHQ, emplNQ Chefen, OHQ, OS, chome1, chome2, Etudes
Habitat en Zus	2	Oui Non

11. Estimation

Les cellules étant définies, il ne nous reste plus qu'à effectuer les estimations. La marche à suivre est la suivante.

Calcul de la moyenne de chaque cellule de VdQ

Pour cela, il suffit d'effectuer un simple croisement entre la variable d'intérêt et la variable de cellulage :

```
proc freq data=iso2; table cell16*isolement04/nofreq nocol nopercnt ; weight poids_i; run;
```

cell16=variable de cellulage ; isolement04=variable d'intérêt. poids_i=poids individuel provenant de Vie de quartier.

Les options nofreq nocol nopercnt permettent de ne récupérer que les données nécessaires.

Calcul de la structure de chaque cellule des 3 Zus de Bretagne

Sous SAS, on obtiendra les effectifs de chaque cellule à partir de la variable de cellulage. Le fichier est limité aux trois Zus de Bretagne sélectionnées grâce à leur numéro.

```
proc freq data=z1;table cell16/nofreq nocol nopercnt;run;
```

L'estimation

L'estimation proprement dite est effectuée avec Excel afin de pouvoir contrôler de visu toutes les cellules. Le tableau 6 représente la feuille de calcul utilisée.

Le coefficient de structure est calculé en rapportant l'effectif d'une cellule des 3 Zus de Bretagne sur le total des effectifs de ces trois Zus.

En rappel :

$\frac{N_{h,z}}{N_z}$ est le coefficient de structure de chaque cellule h dans les 3 Zus de Bretagne z .

L'estimation de l'isolement (colonne 7) s'obtient en faisant le produit ligne à ligne de la moyenne de l'isolement dans chaque cellule de structure définie à partir de Vie de quartier (colonne 3) par la structure sociodémographique du domaine à estimer (colonne 6) .

Tableau 6 - Feuille de calcul Excel; Cellulage version 16 (cell1d)

N° cellule Vie de quartier	Effectifs VdQ	Coefficients de comportement isolement %	N° cellule Zus RP99	Effectif (Nh) du Pt dom	Coeff. structure (Nh/Nz)	Estimation isolement
1	69	5,6	1	323	0,017073686	0,095612644
2	187	6,45	2	235	0,012422032	0,080122106
3	126	23,89	3	415	0,02193678	0,524069669
4	124	12,11	4	549	0,029019981	0,35143197
5	72	6,38	5	186	0,009831906	0,062727561
6	49	9,52	6	258	0,013637805	0,129831906
7	66	9,07	7	202	0,010677661	0,09684639
10	216	23,54			0	0
11	48	14,85	11	814	0,043027804	0,638962892
12	136	9,93	12	274	0,014483561	0,143821757
13	85	18,94	13	323	0,017073686	0,323375621
14	249	19,9	14	533	0,028174226	0,56066709
15	250	13,28			0	0
16	42	29,78	16	471	0,024896924	0,741430384
17	49	18,56	17	76	0,004017338	0,074561793
18	8	0	18	4	0,000211439	0
19	100	19,01	19	111	0,005867428	0,111539803
20	401	17,77	20	408	0,021566762	0,383241357
21	86	4,25	21	272	0,014377841	0,061105825
22	383	2,13			0	0
23	35	2,9	23	274	0,014483561	0,042002326
24	124	11,65	24	406	0,021461042	0,250021144
25	68	12,11	25	111	0,005867428	0,071054551
26	85	9,95	26	286	0,015117877	0,150422878
27	114	25,4	27	151	0,007981816	0,202738133
28	222	17,36			0	0
29	40	23,59	29	669	0,035363146	0,834216619
30	186	6,51			0	0
31	29	32,25	31	294	0,015540755	0,501189343
32	53	10,79	32	38	0,002008669	0,021673538
33	96	17,02	33	239	0,012633471	0,215021672
34	524	6,59			0	0
35	33	6,23	35	868	0,045882229	0,285846284
36	123	23,62	36	198	0,010466223	0,247212179
37	604	15,97	37	782	0,041336293	0,660140607
38	10	6,75			0	0
39	68	2,29	39	124	0,006554604	0,015010043
40	133	5,03	40	88	0,004651655	0,023397822
42	258	4,19			0	0
43	29	21,93	43	832	0,043979279	0,964465588
44	429	5,18			0	0
45	47	11,27	45	567	0,029971456	0,337778306
46	289	4,27	46	491	0,025954118	0,110824083

Tableau 6 - Feuille de calcul Excel; Cellulage version 16 (cell11d)

(suite)

N° cellule Vie de quartier	Effectifs VdQ	Coefficients de comportement isolement %	N° cellule Zus RP99	Effectif (Nh) du Pt dom	Coeff. structure (Nh/Nz)	Estimation isolement
47	128	8,78	47	527	0,027857067	0,244585051
48	172	6,67	48	495	0,026165557	0,174524263
49	292	5,88	49	298	0,015752194	0,092622899
50	294	2,7	50	307	0,016227931	0,043815414
51	176	17,71	51	320	0,016915107	0,29956655
52	448	8,48			0	0
53	36	22,42	53	798	0,042182049	0,945721535
54	717	8,77			0	0
55	46	8,29	55	829	0,0438207	0,363273602
56	359	13,24	56	298	0,015752194	0,208559044
58	235	13,33	58	567	0,029971456	0,399519505
59	48	27,56	59	37	0,001955809	0,053902104
61	442	6,63	61	484	0,0255841	0,169622582
62	220	7,36	62	200	0,010571942	0,077809494
63	100	8,18	63	81	0,004281637	0,035023787
65	68	18,04	65	41	0,002167248	0,039097156
66	322	14,34	66	198	0,010466223	0,150085633
67	160	12,32	67	551	0,0291257	0,358828629
68	36	9,38	68	57	0,003013003	0,028261973
69	95	6,85	69	52	0,002748705	0,018828629
70	180	2,31	70	223	0,011787715	0,027229623
71	39	7,47	71	72	0,003805899	0,028430067
72	74	3,11	72	51	0,002695845	0,008384079
73	164	7,13	73	166	0,008774712	0,062563696
74	55	8,75	74	50	0,002642986	0,023126123
75	113	9,65	75	56	0,002960144	0,028565387
76	168	9,25	76	99	0,005233111	0,04840628
77	68	7,93	77	61	0,003224442	0,025569828
78	110	13,07	78	76	0,004017338	0,052506607
79	46	9,78	79	32	0,001691511	0,016542975
80	46	4,47	80	20	0,001057194	0,004725658
-	11802	-	-	18918	1	13,34

La numérotation est discontinuée à cause des différents regroupements effectués a posteriori. D'autres cellules auraient mérité d'être regroupées au regard de leur effectif réduit mais le regroupement « moyennise » aussi les résultats des autres cellules, les rendant ainsi moins pertinents lors de l'estimation à partir du fichier de structure.

Les différentes estimations petits domaines effectuées à partir des cellulages précédents figurent dans les tableaux suivants. On voit déjà que ce n'est pas le nombre de cellules qui est déterminant dans la qualité de l'estimateur et que les différents choix du cellulage, à partir des mêmes variables, est pour beaucoup dans la qualité des estimations.

Résultats

Après de nombreux essais, l'estimation effectuée la plus proche de la série accuse un écart de 1,0 % avec l'estimateur direct venant des extensions régionales bretonnes de l'enquête. Cette dernière estimation semble suffisamment approchée pour qu'on puisse s'en contenter.

Tableau 7 - Estimations sur les Zus de Rennes et de St Malo

%

	Estimateur direct	Estimateurs Petits domaines							
	(A)	Version 1 cel01		Version 2 cel02		Version 3 cel03		Version 4 cel03	
		55 cellules	Écart à (A)	64 cellules	Écart à (A)	102 cellules	Écart à (A)	85 cellules	Écart à (A)
Ensemble des 3 Zus	14,3	11,0	3,3	11,0	3,3	11,2	3,1	11,2	3,1

Suite

%

	Estimateur direct	Estimateurs Petits domaines							
	(A)	Version seg1		Version seg2		Version 11 cel10b		Version 16 cel11d	
		110 cellules	Écart à (A)	cellules	Écart à (A)	94 cellules	Écart à (A)	75 cellules	Écart à (A)
Ensemble des 3 Zus	14,3	12,1	2,2	12,1	2,2	12,0	2,3	13,3	1,0

12. Variance des estimateurs

L'estimation étant réalisée, on souhaite maintenant connaître la variance qui lui est attachée. Pour cela, il est nécessaire de faire un rappel préalable. La variable d'intérêt Y vaut « 1 » lorsque la condition est remplie, c'est-à-dire quand la personne est isolée et « 0 » dans le cas contraire. On souhaite calculer la variance respective des estimateurs de la moyenne de Y : l'estimateur direct, dit d' « Horwitz-Thomson », et l'estimateur post-stratifié ou « petits domaines ».

Nous nous trouvons dans le cas peu favorable d'un tirage à probabilités inégales, de taille fixe. La pondération (donc les probabilités d'insertion) des individus qui répondent à la question qui nous intéresse (individus KISHS) est obtenue par un tirage sur l'échantillon des ménages, redressé par un calage sur marges (il s'agit des marges du recensement). Ces pondérations au niveau ménage sont en même temps les pondérations individuelles. Il faut les multiplier par les pondérations KISHS pour obtenir le poids final recherché (ou la probabilité finale d'insertion) de l'individu. La probabilité d'insertion des KISHS nous a été pour l'essentiel, fournie par l'unité de méthodologie statistique de l'Insee. Elle provient de l'algorithme programmé des tirages KISHS. Le calcul aboutissant à la probabilité d'insertion dans l'échantillon est assez complexe et parvient à des résultats assez différents suivant les individus. Il faut en outre remarquer que le tirage diffère selon que l'enquête est nationale ou régionale (extension de Bretagne).

De fait, le tirage des ménages de l'extension régionale s'est effectué par la macro-cube, ce qui entraîne un tirage équilibré sur les principales variables sociodémographiques. Pour l'enquête nationale, le tirage des ménages s'est effectué en plusieurs étapes. En revanche, la seconde partie de la pondération (le calcul de la pondération « KISH ») a été sensiblement la même dans les deux cas.

La formule de la variance, utilisée pour calculer la précision des estimateurs du taux d'isolement relationnel, est celle utilisée dans Poulpe (logiciel Insee d'estimation de la précision) pour le cas le plus général du tirage à probabilités inégales et de taille fixe. Celle-ci est mieux adaptée au tirage de l'échantillon régional qu'au tirage de l'échantillon national, lequel est un sondage à plusieurs degrés.

En annexe 9, on établit le calcul d'un estimateur de la variance de notre estimateur de moyenne.

L'estimateur petits Domaines et l'estimateur de sa variance

On a construit un estimateur « petits Domaines » sur les 3 zones urbaines sensibles de Bretagne, pour estimer la moyenne (ou la proportion) de l'indicateur d'isolement relationnel, les cellules h ont été construites suivant la modélisation expliquée précédemment.

L'estimateur de la moyenne des réponses « oui » à la question précédente sur le domaine Z (Z=le groupe des 3 Zus) est :

$$\hat{Y} = \sum_h \frac{N_{Z,h}}{N_Z} \frac{\sum_{k \in S \cap (\text{strateh})} w_k y_k}{\sum_{k \in S \cap (\text{strateh})} w_k} ;$$

Ici, $N_{Z,h}$ représente les effectifs réels de la cellule h dans la zone Z, N_Z les effectifs totaux réels de la zone Z et w_k est exactement : $1/n_k$.

Les coefficients :

$$\frac{\sum_{k \in S \cap (\text{strateh})} w_k y_k}{\sum_{k \in S \cap (\text{strateh})} w_k}$$

sont appelés « coefficients de comportement ». La sommation s'effectue sur l'ensemble de l'échantillon national S, et non pas seulement sur $S_Z = S \cap Z$, car, comme il a été montré précédemment (§ 11), on n'a pas réussi à mettre en évidence une différence de comportement entre la population en ZUS et la population nationale. Ce qui simplifie la construction de cet estimateur « petits domaines ».

Pour calculer la variance sur l'ensemble des cellules, on peut réécrire l'estimateur précédent en introduisant la différence entre y_k et de la moyenne dans la cellule pondérée par le coefficient :

$$Coeff = \frac{N}{N_h} \frac{N_{Z,h}}{N_Z}$$

(ou son estimateur) pour la population totale N et l'effectif de la cellule N_h , qui remplacera les y_k :

Le calcul de la variance se fait par la même formule, en remplaçant y_k par \hat{e}_k :

$$\hat{e}_k = \frac{N}{N_h} \frac{N_{Z,h}}{N_Z} \left(y_k - \frac{\sum_{k \in S \cap (\text{strateh})} w_k y_k}{\sum_{k \in S \cap (\text{strateh})} w_k} \right);$$

ce qui donne :

$$\hat{V}(\hat{Y}) = \frac{1}{N^2} \frac{1}{(1 - \sum_{l \in S} a_l^2)} \left(\sum_{k \in S} (1 - p_k) \frac{\hat{e}_k}{p_k} - A^2 \right);$$

avec:

$$a_l = \frac{(1 - p_l)}{\sum_{k \in S} (1 - p_k)}$$

$$A = \sum_{k \in S} \frac{a_k \hat{e}_k}{p_k} ;$$

Voir la démonstration dans le rapport de Valérie Albouy.

Application numérique variances de Horwitz-Thomson et « petits domaines »

Comme l'échantillon du Blosne, de Maurepas et de St Malo ont été tirés par la macro cube, un tirage équilibré, la formule est bien adaptée à l'estimateur direct, on peut considérer que la valeur de l'écart type est correcte. On retrouve bien une variance qui est de l'ordre de $1/n$, n étant la taille de l'échantillon.

Tableau 8a - Variances des estimateurs directs et post-stratifiés pour les 3 Zus

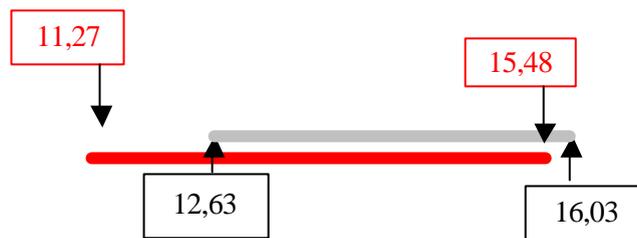
(coefficients de comportement calculés sur France entière)

Seuil de la variable d'isolement	Estimation directe %	Variance	Ecart type %	Coeff. de variation %	Estimateur post-strat. %	Variance	Ecart type %	Coeff. de variation %
1 ^{er} décile	14,33	0.0000715122	0,85	5,9	13,34	0,0001134	1,07	7,98
Demi-médiane	30,97	0.0001373846	1,17	3,8	30,1	0,000235	1,53	11,49
3 ^e décile	36,01	0.0001533566	1,24	3,4	36,2	0,000288	1,70	12,73

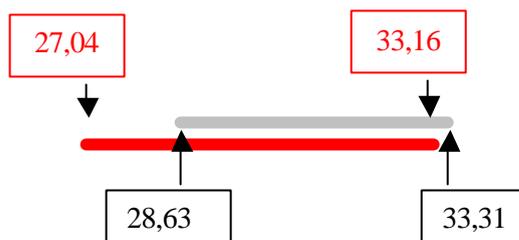
Graphique 1 - Recouvrements des intervalles de confiance pour les 3 Zus de Bretagne

(coefficients de comportement calculés sur la France entière)

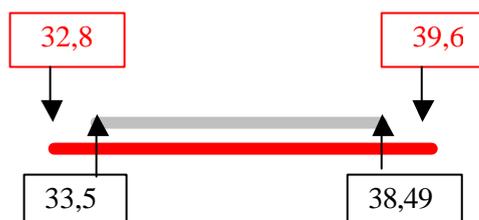
1^{er} décile



2^e décile



3^e décile



— intervalle de confiance de l'estimateur post-stratifié
 — intervalle de confiance de l'estimateur direct

Il y a donc un paradoxe Zus dans ces résultats car on pouvait s'attendre, compte-tenu des 11 500 personnes de l'échantillon national VdQ, à obtenir une précision plus grande que celle de l'estimateur direct portant sur un échantillon de 3 000 personnes. Or, il n'en n'est rien car la structure sociodémographique de l'échantillon national est très différente de celle de nos trois Zus. L'hypothèse initiale avancée pour calculer la variance des estimateurs directs et petits domaines se fondait sur des coefficients de structure proches de 1 pour toutes les cellules. Or, si cela est vraisemblable pour l'estimateur direct (présentant un taux de sondage élevé d'un sixième et bénéficiant de surcroît d'un échantillon équilibré sur des variables définissant aussi les cellules de notre estimateur), il n'en va pas de même pour l'estimateur petits-domaines. En effet, nous comparons pour chaque cellule h la population dans une Zus à la population France entière.

Nous avons aussi testé la construction de l'indicateur à partir des seuls individus résidant en Zus contenus dans l'échantillon national « Vie de quartier ». L'effectif représente 1 203 individus au lieu des 11 500 de l'échantillon total. La proximité des caractéristiques des habitants en Zus a pour conséquence de rapprocher la population de notre domaine avec la population globale que nous utilisons (en l'occurrence les Zus de la France entière). Bien que le nombre d'individus considéré soit évidemment moins important que dans l'enquête nationale le gain de précision (la réduction de l'écart type) sur les Zus est appréciable par cette méthode. Cependant l'estimateur devient particulièrement instable à cause de cellules dégarnies (plusieurs ne contiennent moins de 10 individus), pour cette raison nous nous proposons de conserver l'estimateur calculé à partir de la totalité de l'échantillon national.

Lorsque l'estimation porte sur des zones où les effectifs sont importants, comme Rennes ou Brest, les structures sociodémographiques se rapprochent plus probablement d'une composition moyenne France entière en écrétant les particularités sociodémographiques marquées. De ce fait, l'utilisation des coefficients de comportement basés sur la totalité de l'échantillon national « Vie de quartier » trouve sa pleine justification. Voilà ce que donne la répartition des coefficients de structure lorsqu'on l'applique aux trois Zus de Bretagne sur ceux de France entière :

Calcul des coefficients de structure des trois Zus de Bretagne / France entière

$$(\text{coeff} = \frac{N}{N_h} \frac{N_{zh}}{N_z})$$

Moments			
N	63	Sum Weights	63
Mean	3.45402289	Sum Observations	217.603442
Std Deviation	5.59855248	Variance	31.3437899
Skewness	1.97146113	Kurtosis	2.23843513
Uncorrected SS	2694.92224	Corrected SS	1943.31498
Coeff Variation	162.08788	Std Error Mean	0.70535131

La moyenne est à 3,5 ce qui montre que l'hypothèse d'une structure sociodémographique proche entre les Zus et la population moyenne était fautive.

Calcul des coefficients de structure des trois Zus de Bretagne / Zus de la France entière ;

Moments			
N	64	Sum Weights	64
Mean	1.23905241	Sum Observations	79.2993542
Std Deviation	0.30274933	Variance	0.09165716
Skewness	-0.1399057	Kurtosis	3.5942899
Uncorrected SS	104.030457	Corrected SS	5.77440077
Coeff Variation	24.4339404	Std Error Mean	0.03784367

On voit que la moyenne des coefficients de structure est proche de 1 et que leur dispersion est bien moins grande que dans l'exemple précédent.

Tableau 8b - Variance de l'estimateur post-stratifié pour les 3 Zus (coefficients de comportement calculés sur les individus en Zus)

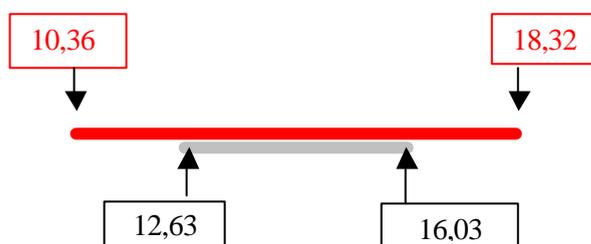
Seuil de la variable d'isolement	Estimation directe %	Variance	Ecart type %	Coeff. de variation %	Estimateur post-strat. %	Variance	Ecart type %	Coeff. de variation %
1 ^{er} décile	14,33	0.0000715122	0,85	5,9	14,34	0.00039	1,98	13,68
Demi-médiane	30,97	0.0001373846	1,17	3,8	32,3	0.00068	2,6	7,98
3 ^e décile	36,01	0.0001533566	1,24	3,4	41,3	0.0008	2,8	6,79

En se restreignant à la population des Zus de l'échantillon national, la variance de l'estimateur post-stratifié ne s'est pas améliorée, au contraire. Les coefficients de structure sont bien moins dispersés mais c'est au détriment du biais puisque l'échantillon, utilisé pour évaluer les coefficients de comportement, passe de 11 500 personnes à un échantillon de 1 200 individus. La conséquence est de retrouver dans certaines cellules relatives aux coefficients de comportement, moins de dix individus !

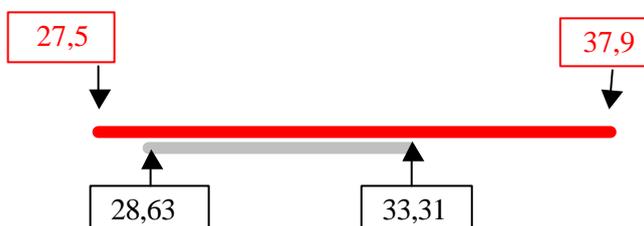
Graphique 2 - Recouvrements des intervalles de confiance pour les 3 Zus

(estimateur post-stratifié construit sur le sous-échantillon de l'ensemble des Zus de vie de quartier)

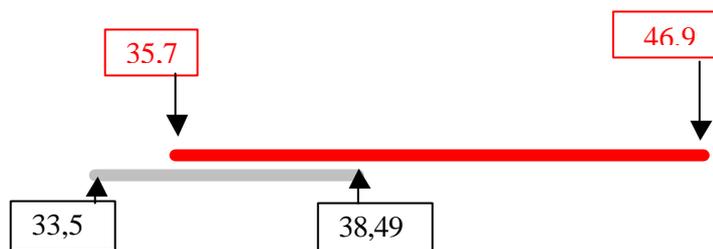
1^{er} décile



2^e décile



3^e décile



— intervalle de confiance de l'estimateur post-stratifié
 — intervalle de confiance de l'estimateur direct

13. Contrôle des estimateurs

On peut aussi douter de la précision et de la stabilité de notre estimateur sur d'autres découpages socio-spatiaux et un seul point de comparaison est insuffisant pour valider l'estimateur post-stratifié puisque le premier point de comparaison sert aussi d'étalonnage de celui-ci. L'Île-de-France représente une strate à part entière dans le plan de sondage de l'enquête, les données bénéficient d'une

représentativité pour cette région. Nous avons donc la possibilité de confronter notre estimateur post-stratifié sur un découpage géographique beaucoup plus important et d'observer son comportement (tableau 9a).

Comportements des indicateurs sur la région Ile-de-France

Les écarts de niveau entre l'estimateur direct et post-stratifié sont réguliers quelque soit le seuil d'isolement retenu, de 11% pour le 1^{er} décile à 13 % pour la demi médiane et le 3^e décile. Cette relative stabilité permet de penser qu'au niveau de la région un coefficient correcteur serait suffisant pour rattraper l'estimation directe. Mais est-ce nécessaire et justifié ? Auparavant, il est nécessaire de vérifier la précision des estimations directe et post-stratifiée (tableau 9b et graphique 3). L'intervalle de confiance de l'estimation directe s'étale de 9,5 % à 14,7 % alors que celui de l'estimation post-stratifiée est réduit de 10,0 % à 11,6 %... Ainsi, l'estimateur direct à 12,1 % est peu précis et il serait dommageable d'aligner l'estimateur petit domaine sur ce niveau d'autant qu'on le ferait sortir des bornes de son intervalle de confiance. L'écart entre la moyenne nationale et l'estimateur petit domaine portant sur l'Île-de-France tend à montrer qu'une zone importante d'estimation moyennise celle-ci. Il est probable que sur un découpage aussi important, les effets de structure se compensent et l'estimateur reflète ce constat.

Les écarts sont bien plus resserrés pour les estimations portant sur les 3 Zus de Bretagne, puisqu'ils varient entre 1 % et 6 % à cause des variations de structure plus marquées. L'application d'un coefficient correcteur ne se justifie pas plus car la valeur de l'estimateur post-stratifié est dans la bande de l'intervalle de confiance des estimations directes.

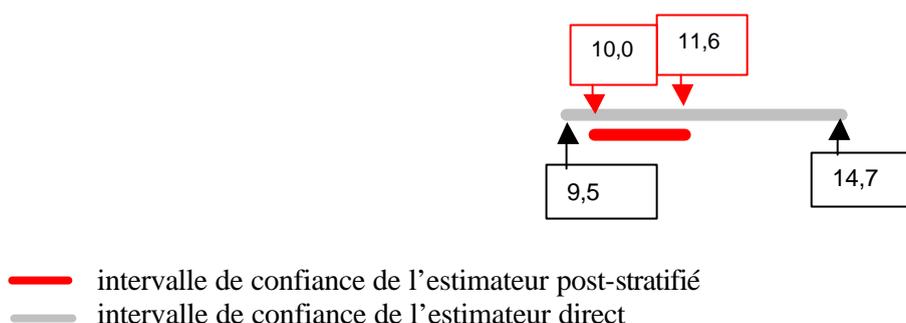
Tableau 9a - Comportement des estimateurs en Île-de-France et sur les 3 Zus

		Estimateurs 1 ^{er} décile %	Indice B/A	Estimateurs 1/2 médiane %	Indice B/A	Estimateurs 3 ^e décile %	Indice B/A
A Direct	IdF	12,1	1,0	30,3	1,0	36,8	1,0
B Post-stratifié		10,8	0,89	26,3	0,87	31,9	0,87
A Direct	3 Zus Breta.	14,3	1,0	31,0	1,0	36,0	1,0
B Post-stratifié		13,3	0,93	30,1	0,97	36,2	1,01

Tableau 9b - Précision de l'indicateur d'isolement en Île-de-France

	Estimateurs 1 ^{er} décile %	Variance	Ecart type %	Coeff. de variation %
Direct	12,1	0.0001589347	1,3	10,4
Post-stratifié	10,8	0,000001671	0,41	3,79

Graphique 3 - Recouvrements des intervalles de confiance pour l'Ile-de-France



Biais possibles

Chacun a par sa propre expérience pu constater des comportements différenciés dans la spontanéité des contacts établis entre le sud, le nord et l'est, ce qui pourrait avoir une influence sur le nombre total des contacts et donc sur le plus ou moins grand nombre d'isolés. Les données de l'enquête n'étant pas représentatives au niveau régional, il est impossible de tester son éventuel effet, notamment en introduisant la variable région dans la régression logistique puisque nous n'aurions pu interpréter les résultats : soit comme un aléa d'échantillonnage, soit un effet purement régional, soit encore les deux mélangés... De plus, il faut souligner que l'effet « régional » est difficile à cerner car il n'y a pas de raison qu'il s'exerce obligatoirement dans les délimitations administratives de la région !

D'autres découpages peuvent interférer notamment les pays, au sens socio-historique du terme, comme a pu l'avancer Le Bras mais sans véritable démonstration (2002), ou d'autres influences socio-géographiques peuvent modifier le comportement moyen régional : le relief donc les obstacles naturels à la communication qui façonnent les individualités et les rapports aux autres, le littoral, le partage d'activités communes qui crée des réseaux de relations plus ou moins formels, etc. Toutefois, cet estimateur est destiné à être interprété en différentiel, son utilisation au sein d'une même région aux caractéristiques géographiques identiques ne présente pas d'inconvénients (tableau 10). Cependant, les estimations réalisées sont proches des estimations directes et les intervalles de confiance des l'estimateur directe recouvre totalement l'estimateur post-stratifié ce qui tendrait à prouver que les biais ne seraient que très faibles.

Tableau 10 - Exemples d'estimations sur les communes de Brest et Rennes

	Estimateur post-strat. 1 ^{er} décile %	Variance	Ecart type %	Coeff. de variation %
Brest	9,73	0,0000109433	0,33	3,39
Rennes	9,36	0,0000145397	0,34	3,54

Vérification de la stabilité des estimateurs d'isolement

Il est possible d'éprouver aussi la stabilité de l'estimateur en fonction des seuils d'isolement retenus. Le niveau de précision étant accepté, il faut alors s'intéresser à la stabilité. Le tableau 11 exprime la plus ou moins grande variabilité des estimateurs en fonction des différents seuils d'isolement retenus. Nous pouvons constater qu'ils ne se comportent pas de façon erratique, notamment sur les grands découpages spatiaux ou socio-spatiaux. Ainsi, les écarts entre premier et 3^e décile sur la France hors Zus et l'Ile-de-France ne dépassent pas respectivement 1 et 4 centièmes de notre indice de variabilité. L'écart maximum sur l'ensemble des Zus se limite à 1,1 dixième. Le différentiel entre les différents estimateurs des 3 Zus de Bretagne atteint 1,9 dixième et 2 dixièmes respectivement pour les estimateurs directs et post-stratifiés, ce qui indique un impact sensible mais cohérent du seuil retenu de l'indicateur sur les découpages plus réduits.

Tableau 11 - Variations des estimateurs selon le seuil retenu

%

	Différents seuils d'isolement					
	1 ^{er} décile		Demi médiane		3 ^e décile	
	%	Indice	%	Indice	%	Indice
Ile de France	12,1	1,12	30,3	1,15	36,8	1,16
Ensemble hors Zus	10,4	0,96	25,7	0,97	30,7	0,97
Ensemble des Zus	15,9	1,47	35,9	1,36	45,5	1,43
3 Zus de Bretagne : direct	14,3	1,32	31,0	1,17	36,0	1,13
3 Zus de Bretagne : post-strat. (Cel11d)	13,4	1,24	30,1	1,14	36,2	1,14
France métropolitaine	10,8	1,00	26,4	1,00	31,8	1,00

Lecture : il y a 12 %, 30,3 % et 36,8 % d'isolés en Ile-de-France, selon le seuil retenu, respectivement du 1^{er} décile, de la demi-médiane ou du 3^e décile. Ce qui rapporté au total représente un indice de 1,12, 1,15 et 1,16. L'indice est calculé en rapportant le % d'isolement de la zone d'estimation sur le % national.

Source : Vie de quartier, Insee.

18. En conclusion

Par nature, les estimateurs post-stratifiés sont des estimateurs structurels, c'est-à-dire qu'ils fournissent le niveau « attendu » d'un indicateur en fonction de la seule composition sociodémographique du domaine étudié. Cet estimateur ne fait que moduler, à un niveau fin et en fonction des variables « explicatives » liées au phénomène, ici l'isolement relationnel, l'estimation directe nationale. La nature structurelle de l'estimateur post-stratifié utilisé dans cette étude montre que la bonne réussite de l'estimation est conditionnée par la plus grande force déterministe du comportement à estimer.

D'un point de vue instrumental, deux éléments principaux interviennent dans la plus ou moins grande qualité d'un estimateur post-stratifié : la bonne définition des variables actives et le découpage fin du fichier de données de la variable d'étude. Comme on a pu le voir, ce découpage est contraint par la taille du fichier, ce qui entraîne des arbitrages entre le nombre de modalités et de variables introduites, ce qui est dommageable à la précision de l'estimation finale. De plus, il n'est pas tenu compte des possibles effets régionaux, des particularités socio-géographiques ou des aléas dus à l'hétérogénéité individuelle ou aux variables inobservées. Ces remarques montrent à l'évidence que l'estimateur post-stratifié ne peut se substituer véritablement à un estimateur direct et que leur statut respectif n'est pas identique. A cet égard, l'utilisation de cet estimateur dans une publication

Néanmoins, cet estimateur post-stratifié répond non seulement à l'ambition initiale qui était de produire une estimation plus précise que la seule moyenne nationale mais fournit également des estimations qui ne sont pas sans intérêt sur un zonage beaucoup plus restreint. Il reste qu'une zone d'estimation restreinte comme celle de nos 3 Zus de Bretagne représente un effectif insuffisant (près de 19 000 individus) pour obtenir un écart type autorisant des comparaisons fiables entre différentes zones d'étude mais il pourrait être objecté que, dans une mesure à peine moindre, c'est également un reproche à adresser aux estimateurs directs portant pourtant sur un échantillon régional d'environ 3 000 individus...

Pour finir, la qualité de la définition de la structure dont dépend aussi la qualité de l'estimation s'appuie sur le recensement de 1999. Au fur et à mesure qu'on s'en éloigne, la source devient moins pertinente et les estimations qui en dépendent aussi. Il est alors nécessaire d'adosser les estimations sur un fichier plus récent surtout si la structure sociodémographique du domaine tend à se modifier sensiblement au fil des ans.

Bibliographie méthodologique

- Albouy V., Rapport de stage Ensaë, Petits domaines et HID, 2000.
- Ardilly P., « Les techniques de sondage », Technip, 1994.
- Attal K., Sautory O., « Estimations de données régionales à l'aide de techniques d'analyse multidimensionnelle », Document de travail UMS n° 9807.
- Couet C., « Estimation locales dans le cadre de l'enquête HID », document de travail de la DSDS, n° F0207, Insee.
- Destandeu S., « Estimation sur des petits domaines. Application à l'enquête Education », Insee-méthodes n° 69-70-71, Insee 1996.
- Deville J.-C., « Estimation de variance pour des statistiques complexes : Techniques des résidus et de linéarisation », Document de travail, n° 9802, Insee.
- Dumontier F., de Singly F., Thélot C., « La lecture moins attractive qu'il y a vingt ans ? », Économie et statistique n° 233, juin 1990.
- Dumontier F., Guillaumat-tailliet F., « Les actifs patrimoniaux en 1996. Analyses systématiques des taux de détention. Tome 1 : dossier de segmentations », Document de travail, n° F9614, Août 1996.
- Jaeger F., « Méthode d'utilisation d'enquête à un niveau géographique où l'échantillon est faible », Insee-méthodes n° 29-30-31, 1992.
- Gelein B., « AFM et segmentation : application à une typologie de zones d'emploi en fonction du taux de chômage : module analyse des données approfondie », mémoire professionnel FCDA, Polycopié, Bruz, ENSAI, 2002.
- Godechot O., « Plus d'amis, plus proches. Essai de comparaison de 2 enquêtes peu comparables », document de travail à lucarne de l'UMS, n° 0004, Insee, 2000.
- Hidiroglou M. A., « Estimation pour les petits domaines : théorie et pratique à Statistique Canada », Insee-méthodes n° 29-30-31, 1992.
- Leblanc D., Lollivier S., Marpsat M., Verger D., « L'économétrie et l'étude des comportements », document de travail, n° 0001, Insee.
- Marpsat M., « La modélisation des « effets de quartier » aux États-unis », Population, Ined, 1999.
- Marpsat M., Laurent R., « Le chômage des jeunes est-il aggravé par l'appartenance à un quartier en difficulté ? », in « Ces quartiers dont on parle », Éditions de l'aube, 1997.
- Saporta G., « Analyse des données approfondies : segmentation », Polycopié, Bruz : ENSAI, 2000.
- Singh M.P., Gambino J., Mantel H.J., « Les petites régions : problèmes et solutions », Techniques d'enquête, Vol. 20 n°1, Statistique Canada, 1994.
- Tillé Y., « Théorie des sondages », *Sciences sup*, Dunod, 2001.
- Documents de la session 4 des journées de méthodologie statistique du 11 et 12 décembre 1996.
- Document de travail à lucarne, « Estimation locales dans le cadre de l'enquête HID. Démarche suivie », en cours de préparation, 2003.

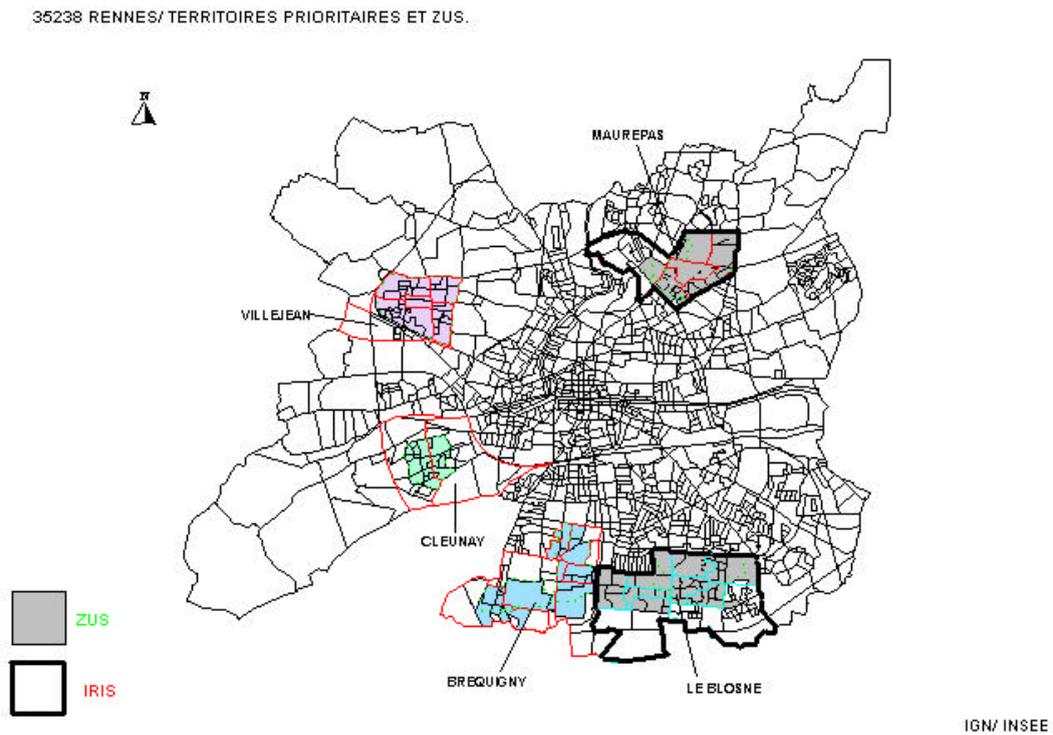
Bibliographie sur la sociabilité et l'isolement

- Blanpain N., Pan Ké Shon J.-L., « La sociabilité des personnes âgées », *Insee-première*, n° 644, mai 1999a
- Blanpain N., Pan Ké Shon J.-L., « A chaque étape de la vie, ses relations », *Données sociales*, 1999b.
- Bourdieu P., « Le capital social », *Actes de la recherche en sciences sociales*, n° 31, janvier 1980a.
- Bourdieu P., « Le sens pratique », Paris, Éd. de Minuit, 1980b.
- Bozon M., « Vie quotidienne et rapports sociaux dans une petite ville de province. La mise en scène des différences », Lyon, Presses universitaires de Lyon, 1984.
- Castel R., « Les métamorphoses de la question sociale », Paris, Fayard, 1995.

- Choquet O., « Les sorties une occasion de contacts », *Économie et statistique*, Insee, 19 ??.
- Crenner E. « La parenté : un réseau de sociabilité actif mais concentré », *Insee première*, n° 600, 1998.
- Degenne A., Fournier I., Marry C., Mounier L., « Les relations sociales au cœur du marché du travail », *Sociétés contemporaines*, n° 5, 1991.
- Dickes P., « Pauvreté en termes de conditions d'existence », recherche menée pour le compte de la MIRE, université de Nancy II, Adeps, 1992.
- Ehrenberg A., « La fatigue d'être soi : dépression et société », Paris, Odile Jacob, 1998.
- Ferrand A., Mounier L., « L'échange des paroles sur la sexualité : une analyse des relations de confiance. » - In : *Sexualité et sciences sociales : les apports d'une enquête*, numéro coordonné par M. Bozon et H. Leridon, *Population*. - N° 5, p. 1451-1476, 1993
- Forsé M., « Capital social et emploi », *L'Année sociologique*, vol.47-1, 1997.
- Forsé M., « Rôle spécifique et croissance du capital social », *Revue de l'OFCE*, n° 76, 2001.
- Galland O., « *Vie solitaire et solitude, le cas des jeunes* », *L'Année sociologique*, 1993 - vol. 43.
- Giret J.-F., Karaa A., Plassard J.-M., « Mode d'accès à l'emploi des jeunes et salaires », *Formation Emploi*, N° 54, 1996.
- Granovetter M. S., « The strength of weak ties », *American journal of sociologie*, vol. 78, n° 6, 1973.
- Hannoun M., « Nos solitudes. Enquête sur un sentiment », *L'épreuve des faits*, Seuil, 1991.
- Héran F., « La sociabilité, une pratique culturelle », *Économie et Statistique*, n° 216, Insee, décembre 1988.
- Héran F. « Trouver à qui parler : le sexe et l'âge de nos interlocuteurs », *Données sociales*, Insee, 1990.
- Kaufmann J.-C., « L'état des savoirs », *Rapport pour la communauté européenne*, Eurostat, 1994.
- Kaufmann J.-C., « Vie hors couple, isolement et lien social : figures de l'inscription relationnelle », *Revue française de sociologie*, n° XXXV, 1994.
- Kaufmann J.-C., « *Trois contextes sociaux de l'isolement* », *Revue française des Affaires sociales*, 1994.
- Le Bras H., « Une autre France », Odile Jacob, 2002.
- Lemel Y., Paradeise C., « La sociabilité », les comportements de sociabilité d'après l'enquête Loisirs, document à lucarne, Insee, 1967.
- Marpsat M., « Les échanges au sein de la famille. Héritages, aides financières, garde des enfants et visites aux grands-parents », *Économie et statistique*, n° 239, 1991.
- Marry C., « Origine sociale et réseaux d'insertion des jeunes ouvriers », *Formation-Emploi*, n° 4, 1983.
- Pan Ké Shon J.-L., « Isolement relationnel et mal-être », *Insee première* n° 931, novembre 2003.
- Pan Ké Shon J.-L., « Être seul », *Données sociales*, novembre 2002.
- Pan Ké Shon J.-L., « Vivre seul, sentiment de solitude et isolement relationnel », *Insee première* n° 678, 1999.
- Rivière C.-A. « Le téléphone : un facteur d'intégration sociale », *Économie et Statistique*, n° 216, Insee, novembre 2001.
- Wolff F.-C., « Transferts monétaires inter vivos et cycle de vie », *Revue économique*, Vol. 51, 2000.
- Wolff F.-C., « Les transferts versés aux enfants et aux parents : altruisme ou échange intertemporel ? », *Économie et prévision*, n° 142, 2000.
- « Mesurer la pauvreté aujourd'hui », *Économie et Statistique*, n° 308-309-310, Insee, 1997.

ANNEXES

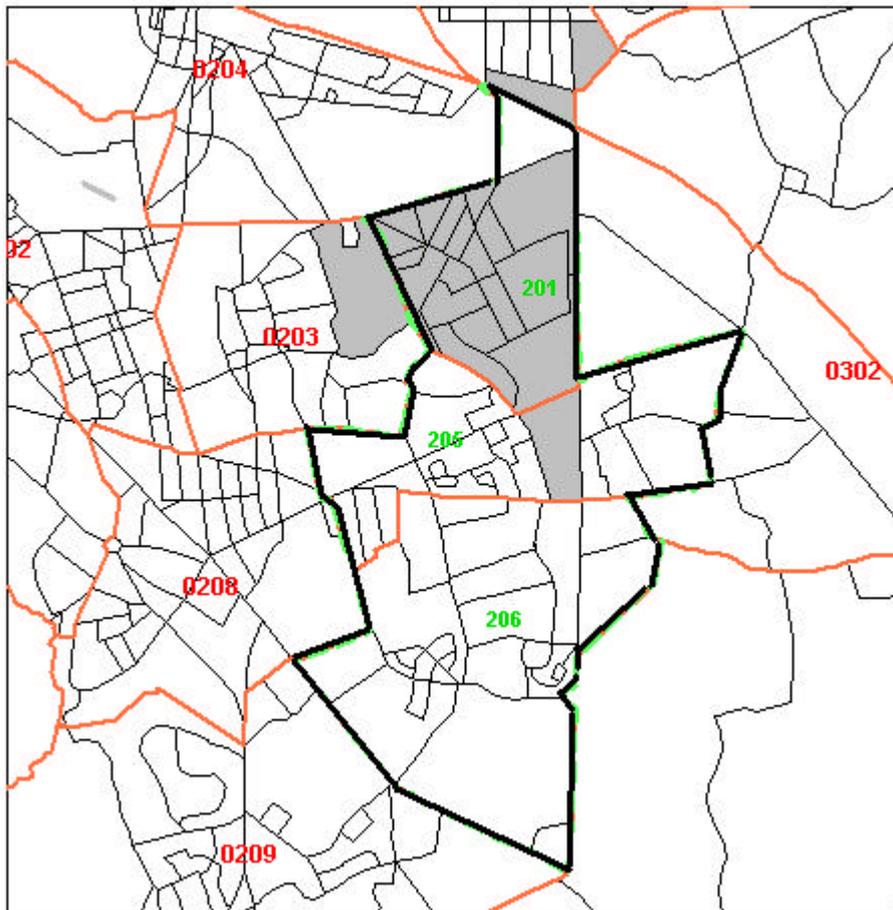
Annexe 1 - Quartiers Le Blossne et Maurepas à Rennes



Localisation des quartiers (respectivement en grisé au nord et au sud).

La zone grisée correspond au découpage exact des zones urbaines sensibles de Maurepas et de Le Blossne, la délimitation en gras correspond au périmètre des iris de ces quartiers contenant au moins une part de ces Zus.

Annexe 2 - Quartier La Découverte à Saint-Malo



En trait gras, les 3 iris formant le quartier prioritaire de La Découverte de St Malo (201, 205, 206)
en grisé, la partie Zus de la Découverte
(pour PCV, on n'a que la partie grisée compris dans le périmètre vert)

Annexe 3

Modèle isolement France entière

The LOGISTIC Procedure

Model Information

Data Set	WORK.IS01
Response Variable	isolement04
Number of Response Levels	2
Number of Observations	11802
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	isolement04	Total Frequency
1	1	1311
2	0	10491

NOTE: 117 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept	
	Only	and Covariates
AIC	8234.446	7736.616
SC	8241.822	7972.649
-2 Log L	8232.446	7672.616

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > Chi Sq
Likelihood Ratio	559.8302	31	<.0001
Score	565.2894	31	<.0001
Wald	508.7069	31	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > Chi Sq
Intercept	1	-2.1156	0.0810	682.2936	<.0001
a1519	1	-0.9335	0.2169	18.5155	<.0001
a2025	1	-0.6087	0.1381	19.4270	<.0001
a2534	1	-0.3870	0.1025	14.2559	0.0002
a7099	1	0.4908	0.1028	22.7876	<.0001
dipl o0	1	0.3793	0.0839	20.4499	<.0001
dipl o1	1	0.1834	0.0830	4.8824	0.0271
dipl o4	1	-0.1753	0.1232	2.0244	0.1548
seuve	1	-0.1416	0.1154	1.5050	0.2199
couss	1	-0.2150	0.0829	6.7262	0.0095
couen	1	-0.3146	0.2031	2.3988	0.1214
mopar	1	0.2108	0.1090	3.7384	0.0532
sixplus	1	0.4412	0.1560	7.9942	0.0047
eAfric	1	0.4998	0.1105	20.4499	<.0001
relinon	1	0.2864	0.0688	17.3538	<.0001
reli ref	1	0.1828	0.1508	1.4692	0.2255
reli reg	1	-0.2117	0.0977	4.6893	0.0304
Asocivic	1	-0.3462	0.1303	7.0640	0.0079
asoancie	1	-0.4822	0.1563	9.5245	0.0020
asosport	1	-0.5240	0.1176	19.8489	<.0001
asocul tu	1	-0.5835	0.1354	18.5615	<.0001

xDi ml a2	1	-0.2282	0.0812	7.8899	0.0050
xDi m3a4	1	-0.2665	0.0812	10.7793	0.0010
pcs03b	1	-0.3176	0.1057	9.0289	0.0027
pcs02b	1	0.1942	0.0912	4.5322	0.0333
uc00a20	1	0.5338	0.2097	6.4782	0.0109
uc20a30	1	0.3743	0.1542	5.8930	0.0152
uc30a50	1	0.3311	0.0833	15.8083	<.0001
ucsuper	1	0.2935	0.1599	3.3699	0.0664
ci teaut	1	0.1348	0.0920	2.1469	0.1429
handperm	1	0.4074	0.1505	7.3217	0.0068
xzus	1	0.2422	0.1046	5.3637	0.0206

Association of Predicted Probabilities and Observed Responses

Percent Concordant	68.9	Somers' D	0.387
Percent Discordant	30.2	Gamma	0.391
Percent Tied	0.8	Tau- a	0.076
Pairs	13753701	c	0.694

Linear Hypotheses Testing Results

Label	Chi - Square	DF	Pr > Chi Sq
assos	63.1540	4	<.0001
age	63.0623	4	<.0001
rel ig	29.0237	3	<.0001
di plo	25.7316	3	<.0001
revenUC	24.7495	4	<.0001
Pnai ss	20.4499	1	<.0001
pcsb	16.3780	2	0.0003
menage	14.8466	4	0.0050
axe1	14.5905	2	0.0007
famnb	7.9942	1	0.0047
handi cap	7.3217	1	0.0068
habi typ	2.1469	1	0.1429

Annexe 4

Modèle contracté

The LOGISTIC Procedure

Model Information	
Data Set	WORK.IS01
Response Variable	isolement04
Number of Response Levels	2
Number of Observations	11802
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile		
Ordered Value	isolement04	Total Frequency
1	1	1311
2	0	10491

NOTE: 117 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.
Model Fit Statistics

Criterion	Intercept and Covariates	
	Intercept Only	Intercept and Covariates
AIC	8234.446	7857.375
SC	8241.822	8004.896
-2 Log L	8232.446	7817.375

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > Chi Sq
Likelihood Ratio	415.0709	19	<.0001
Score	426.1344	19	<.0001
Wald	388.8288	19	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard		
			Error	Chi-Square	Pr > Chi Sq
Intercept	1	-2.1434	0.0667	1032.7620	<.0001
a1519	1	-0.7667	0.2167	12.5241	0.0004
a2025	1	-0.4144	0.1337	9.6046	0.0019
a2534	1	-0.3225	0.1012	10.1666	0.0014
a7099	1	0.4779	0.1005	22.6278	<.0001
dipl o0	1	0.5555	0.0803	47.8273	<.0001
dipl o1	1	0.2676	0.0812	10.8495	0.0010
dipl o4	1	-0.2420	0.1224	3.9105	0.0480
pcs02b	1	0.0798	0.0885	0.8132	0.3672
pcs03b	1	-0.2885	0.1055	7.4807	0.0062
seuve	1	-0.1866	0.1139	2.6839	0.1014
couss	1	-0.2483	0.0818	9.2052	0.0024
couen	1	-0.4165	0.2029	4.2162	0.0400
mopar	1	0.2879	0.1069	7.2533	0.0071
sixplus	1	0.5222	0.1512	11.9277	0.0006
eAfric	1	0.4786	0.1065	20.1863	<.0001
xDim1a2	1	-0.1791	0.0799	5.0266	0.0250
xDim3a4	1	-0.2346	0.0801	8.5778	0.0034
citeaut	1	0.1710	0.0914	3.5029	0.0613
xzus	1	0.2603	0.1041	6.2516	0.0124

Association of Predicted Probabilities and Observed Responses

Percent Concordant	66.1	Somers' D	0.337
Percent Discordant	32.4	Gamma	0.343
Percent Tied	1.6	Tau-a	0.067
Pairs	13753701	c	0.669

Linear Hypotheses Testing Results

Label	Chi-Square	DF	Pr > Chi Sq
age	48.8406	4	<.0001
diplo	60.0882	3	<.0001
Pnaiss	20.1863	1	<.0001
pcsb	9.5091	2	0.0086
menage	25.0825	4	<.0001
axe1	10.8058	2	0.0045
famnbe	11.9277	1	0.0006
habityp	3.5029	1	0.0613

Extrait du programme sas pour hiérarchiser les variables

```
proc logistic descending data=Isol ;
model   isolement04=
        a1519 a2025 a2534      a7099
        diplo0 diplo1        diplo4
        xpcs02 xpcs03
        ymonop ymena2
        sixplus eAfric
        xDim1a2 xDim3a4
        citeaut xzus ;

age: test a1519=0, a2025=0, a2534=0, a7099=0 ;
diplo: test diplo0=0, diplo1=0, diplo4=0 ;
menage: test ymonop=0, ymena2=0;
Pnaiss: test eAfric=0;
axe1: test xDim1a2=0, xDim3a4=0 ;
famnbe: test sixplus=0;
pcs: test xpcs02=0, xpcs03=0;
habityp: test citeaut=0;
zus: test xzus=0;

run;
```

Annexe 5

Modèle pour test de la significativité de la différence Zus et non Zus

The LOGISTIC Procedure

Model Information

Data Set	WORK. R1
Response Variable	isolement04
Number of Response Levels	2
Number of Observations	11802
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	isolement04	Total Frequency
1	1	1311
2	0	10491

NOTE: 117 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept and Covariates	
	Intercept Only	Intercept and Covariates
AIC	8234.446	7763.181
SC	8241.822	8220.494
-2 Log L	8232.446	7639.181

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > Chi Sq
Likelihood Ratio	593.2658	61	<.0001
Score	610.5476	61	<.0001
Wald	539.5161	61	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard		Chi-Square	Pr > Chi Sq
		Estimate	Error		
Intercept	1	-2.3611	0.1121	443.4707	<.0001
xzus	1	0.0345	0.3309	0.0109	0.9170
a1519	1	-1.0508	0.2423	18.8122	<.0001
a2025	1	-0.7000	0.1562	20.0803	<.0001
a2534	1	-0.3866	0.1146	11.3755	0.0007
a7099	1	0.5051	0.1048	23.2461	<.0001
dipl0	1	0.3873	0.0918	17.7869	<.0001
dipl01	1	0.1179	0.0892	1.7459	0.1864
dipl04	1	-0.1801	0.1283	1.9715	0.1603
seuve	1	-0.1012	0.1210	0.6989	0.4032
couss	1	-0.2214	0.0874	6.4225	0.0113
couen	1	-0.1988	0.2276	0.7629	0.3824
mopar	1	0.2298	0.1227	3.5077	0.0611
sixplus	1	0.4794	0.1847	6.7365	0.0094
eAfri c	1	0.5172	0.1305	15.7148	<.0001
pcs01	1	0.4572	0.1097	17.3784	<.0001
pcs02	1	0.2822	0.0974	8.3855	0.0038
relinon	1	0.2326	0.0740	9.8839	0.0017

reli ref	1	0.1269	0.1637	0.6006	0.4383
reli reg	1	-0.2122	0.1064	3.9783	0.0461
Asoci vic	1	-0.2989	0.1345	4.9373	0.0263
asoancie	1	-0.4724	0.1592	8.8108	0.0030
asosport	1	-0.5412	0.1226	19.4856	<.0001
asocul tu	1	-0.5714	0.1410	16.4130	<.0001
xDim1a2	1	-0.2215	0.0898	6.0877	0.0136
xDim3a4	1	-0.2090	0.0884	5.5884	0.0181
uc00a20	1	0.6174	0.2321	7.0783	0.0078
uc20a30	1	0.4992	0.1755	8.0883	0.0045
uc30a50	1	0.3502	0.0927	14.2854	0.0002
ucsuper	1	0.2608	0.1619	2.5949	0.1072
ci teaut	1	0.0989	0.1105	0.8007	0.3709
handperm	1	0.3716	0.1708	4.7299	0.0296
za1519	1	-0.7912	0.4444	3.1703	0.0750
za2025	1	-0.1450	0.3187	0.2072	0.6490
za2534	1	-0.2745	0.2391	1.3185	0.2509
za7099	1	0.1698	0.3673	0.2137	0.6438
zdi pl o0	1	0.6114	0.2235	7.4870	0.0062
zdi pl o1	1	0.7658	0.2387	10.2959	0.0013
zdi pl o4	1	-0.1228	0.4709	0.0680	0.7942
zseuve	1	-0.5623	0.4097	1.8831	0.1700
zcouss	1	-0.1001	0.2566	0.1523	0.6964
zcouen	1	-0.8938	0.4677	3.6528	0.0560
zmopar	1	0.1944	0.2438	0.6361	0.4251
zsi xplus	1	0.5454	0.3075	3.1462	0.0761
zAfri c	1	0.5996	0.2180	7.5686	0.0059
zpcs01	1	0.1580	0.2474	0.4080	0.5230
zpcs02	1	0.0566	0.2081	0.0740	0.7856
zrelinon	1	0.7230	0.1956	13.6602	0.0002
zreli ref	1	0.6046	0.4038	2.2425	0.1343
zreli reg	1	-0.1826	0.2540	0.5170	0.4721
zAsoci vic	1	-1.0570	0.5419	3.8047	0.0511
zasoancie	1	-0.6040	0.7803	0.5991	0.4389
zasosport	1	-0.3888	0.4327	0.8072	0.3689
zasocul tu	1	-0.8708	0.4987	3.0485	0.0808
zxDim1a2	1	-0.3418	0.2022	2.8582	0.0909
zxDim3a4	1	-0.5971	0.2089	8.1691	0.0043
zuc00a20	1	0.1811	0.5170	0.1228	0.7260
zuc20a30	1	0.1004	0.3479	0.0832	0.7729
zuc30a50	1	0.2229	0.2076	1.1524	0.2830
zucsuper	1	0.8584	1.2180	0.4966	0.4810
zci teaut	1	0.3344	0.1804	3.4374	0.0637
zhandperm	1	0.8685	0.3450	6.3355	0.0118

Association of Predicted Probabilities and Observed Responses

Percent Concordant	69.3	Somers' D	0.394
Percent Discordant	29.9	Gamma	0.398
Percent Tied	0.8	Tau-a	0.078
Pairs	13753701	c	0.697

La lettre z en début du nom des variables désigne que celle-ci est en Zus

Suite

Significativité des différences

Linear Hypotheses Testing Results

Label	Wald Chi - Square	DF	Pr > Chi Sq
age	0.2631	1	0.6080
Test 2	2.4452	1	0.1179
Test 3	0.1786	1	0.6726
Test 4	0.7707	1	0.3800
dipl o	0.8608	1	0.3535
Test 6	6.4653	1	0.0110
Test 7	0.0138	1	0.9066
menage	1.1648	1	0.2805
Test 9	0.2000	1	0.6547
Test 10	1.7854	1	0.1815
Test 11	0.0168	1	0.8968
famibe	0.0339	1	0.8538
Pnais s	0.1054	1	0.7455
pcs	1.2225	1	0.2689
Test 15	0.9636	1	0.3263
relig	5.4972	1	0.0190
Test 17	1.2024	1	0.2728
Test 18	0.0116	1	0.9144
assos	1.8438	1	0.1745
Test 20	0.0273	1	0.8688
Test 21	0.1148	1	0.7347
Test 22	0.3338	1	0.5635
axe1	0.2961	1	0.5863
Test 24	2.9278	1	0.0871
revenUC	0.5928	1	0.4414
Test 26	1.0473	1	0.3061
Test 27	0.3136	1	0.5755
Test 28	0.2365	1	0.6267
habi tat	1.2396	1	0.2655
handi cap	1.6657	1	0.1968

Suite - Programme sas

Logit pour test de la significativité de la différence

```
proc logistic descending data=r1 ;
model   isolement04= xzus
        a1519 a2025 a2534          a7099
        dipl o0 dipl o1      dipl o4
        seuve couss couen mopar
        sixplus eAfric
        pcs01 pcs02
        relinon reliref relireg Asocivic asoancie asosport asocul tu
        xDi m1a2 xDi m3a4
        uc00a20 uc20a30 uc30a50          ucsuper
        citeaut handperm

        za1519 za2025 za2534          za7099
        zdipl o0 zdipl o1      zdipl o4
        zseuve zcouss zcouen zmopar
        zsixplus zAfric
        zpcs01 zpcs02
        zrelinon zreliref zrelireg zAsocivic zasoancie zasosport zasocul tu
        zxDi m1a2 zxDi m3a4
        zuc00a20 zuc20a30 zuc30a50      zucsuper
        zciteaut zhandperm ;

age: test a1519-za1519=0;
test a2025-za2025=0;
test a2534-za2534=0 ;
test a7099-za7099=0 ;

assos: test asosport -zasosport=0;
test asoancie -zasoancie=0;
test asoci vic -zasoci vic=0;

relig: test relinon- zrelinon=0 ;
test reliref- zreliref=0 ;
test relireg- zrelireg=0;

dipl o: test dipl o0- zdipl o0=0;
test dipl o1- zdipl o1=0;
test dipl o4- zdipl o4=0;

revenUC: test uc00a20- zuc00a20=0;
test uc20a30- zuc20a30=0;
test uc30a50- zuc30a50=0;
test ucsuper- zucsuper=0;

Pnaiss: test eAfric- zafri c=0;

menage: test seuve- zseuve=0;
test couss- zcouss=0;
test couen- zcouen=0;
test mopar- zmopar=0;

axe1: test xDi m1a2- zxdi m1a2=0;
test xDi m3a4- zxdi m3a4=0 ;

famnb: test sixplus- zsixplus=0;

habitat: test citeaut- zciteaut=0;
pcs: test pcs01- zpcs01=0;
test pcs02- zpcs02=0;

run;
```

Annexe 6

Différentes versions du cellulaire

Version 2

Variables	Nbre de modalités	Modalités
Diplômes	4	Sans CEP, BEPC CAP, BAC, >bac+2 Bac +2
Age	4	15 à 19 ans 20 à 34 ans 35 à 69 ans 70 et plus
Type de ménage	3	Veuf, couple sans enfant, enfant d'un couple Monoparent adulte Autres cas
Pays de naissance	2	Africain Autres cas
Ménage de 6 personnes ou plus	2	Oui Non

Version 3

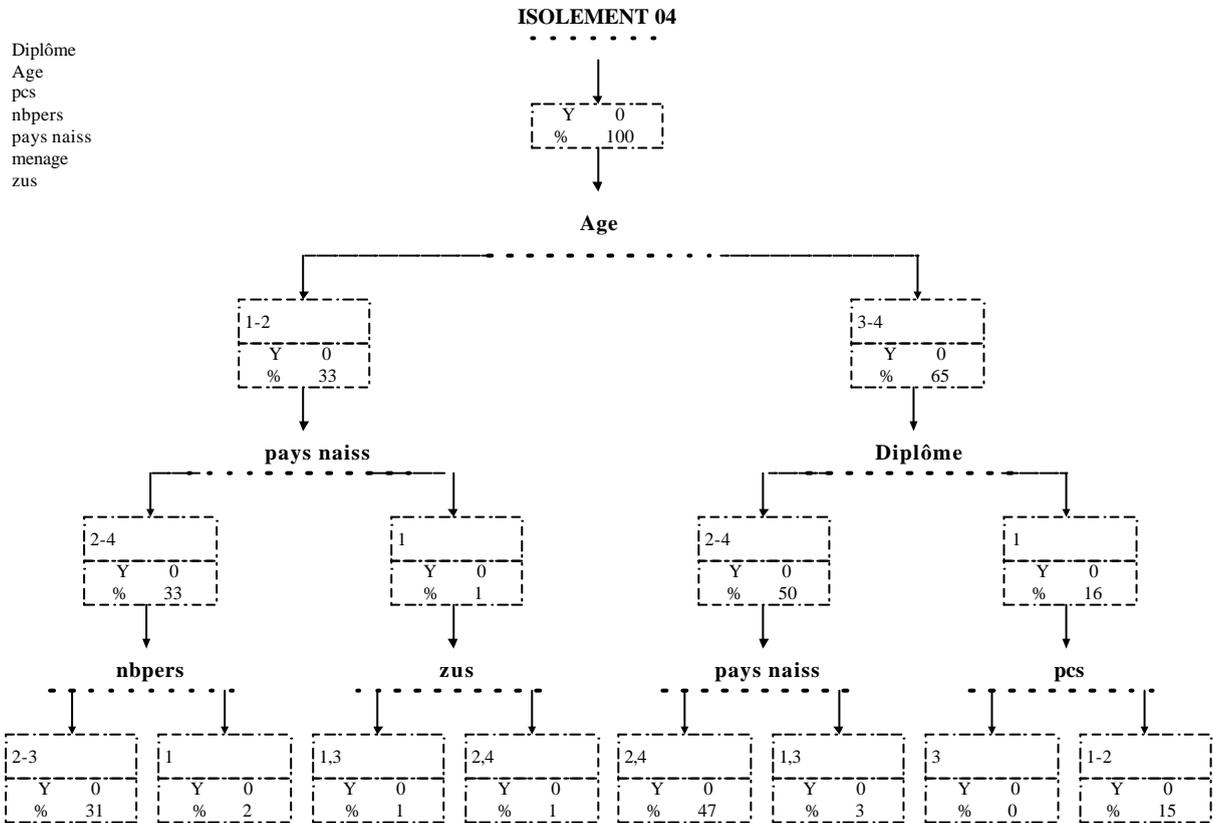
Variables	Nbre de modalités	Modalités
Diplômes	4	Sans CEP, BEPC CAP, BAC, >bac+2 Bac +2
Age	4	15 à 19 ans 20 à 34 ans 35 à 69 ans 70 et plus
Type de ménage	3	Célibataire, adulte d'un cple avec enfant Monoparental Veuf, séparé, couple sans enfant, enfant d'un couple, autre ménage
Pays de naissance	2	Africain Autres cas
Ménage de 6 personnes ou plus	2	Oui Non
PCS regroupée	3	Agric, artcom, AutrCS, retraité Cadres, profIn, emplHQ, emplNQ Chefen, OHQ, OS, chome1, chome2, Etudes

Le fractionnement des cellules à partir de la dernière variable n'est appliquée qu'aux cellules ayant un effectif supérieur à 200.

Version 4

Variables	Nbre de modalités	Modalités
Diplômes	4	Sans CEP, BEPC CAP, BAC, >bac+2 Bac +2
Age	4	15 à 19 ans 20 à 34 ans 35 à 69 ans 70 et plus
Type de ménage	3	Célibataire, adulte d'un couple avec enfant Monoparental Veuf, séparé, couple sans enfant, enfant d'un couple, autre ménage
Pays de naissance	2	Africain Autres cas
Ménage de 6 personnes ou plus	2	Oui Non
PCS regroupée	3	Agric, artcom, AutrCS, retraité Cadres, profIn, emplHQ, emplNQ Chefen, OHQ, OS, chome1, chome2, Etudes

Annexe 7 - segmentation



Annexe 7 - suite

« La macro SEGMENT réalise une segmentation de une ou plusieurs variables sur un certain nombre de critères explicatifs en utilisant un principe de décomposition au vu d'une stratification en deux sous populations rendant maximale la part de variance expliquée. Les variables expliquées ou explicatives doivent être introduites sous forme numérique. Les variables expliquées sont quantitatives (éventuellement dichotomiques). Les variables explicatives sont discrètes, leurs modalités étant comprises entre 1 et 50. Les différentes modalités ne sont pas obligées de se suivre. »

Y=1 correspond à la variable d'intérêt : être isolé

Y=0 correspond à la variable d'intérêt : être non-isolé

C'est cette dernière qui est ici estimée

Les % indiqués sont ceux de la population totale représentée dans la modalité.

Annexe 8

Exemple de Programme de calcul de la variance de l'estimateur post-stratifié (programme sur 4 strates seulement, construites à partir de la variable AGE)

Rappel de la formule faisant l'objet du programme.

$$\hat{V}(\hat{Y}) = \frac{1}{N^2} \frac{1}{(1 - \sum_{l \in S} a_l^2)} \left(\sum_{k \in S} (1 - p_k) \left(\frac{\hat{e}_k}{p_k} - A^2 \right) \right);$$

et toujours :

$$a_l = \frac{(1 - p_l)}{\sum_{k \in S} (1 - p_k)}$$

$$A = \sum_{k \in S} \frac{a_k \hat{e}_k}{p_k};$$

```
data fichier_pour_calcul_variance ;
    merge indiv_echant(in=un) estimateur(in=deux)
          strates_zus_4strates(in=trois); by ZUS ;
    if un and deux and trois;
    run ;

data estim_ek ;set fichier_pour_calcul_variance ;
    if ( strate_age_E=1 )then do ; EK =NZ1 * (ELEOUI - (insc1 / wz1) )/ NZT1 ;
    end ;
    if ( strate_age_E=2 ) then do; EK =NZ2 * (ELEOUI - (insc2 / wz2) )/ NZT2 ;
    end ;
    if ( strate_age_E=3)then do ; EK =NZ3 * (ELEOUI - (insc3 / wz3) )/ NZT3 ;
    end ;
if ( strate_age_E=4)then do ; EK =NZ4 * (ELEOUI - (insc4 / wz4) )/ NZT4 ;
    end ;
    ZUS="5304032" ;
    EK=EK * NTOT/NT ;
    run ;

/* calcul des totaux des sommes de  $1 - p_k$  */
data tot_al ;set estim_ek end=fini;
if _N_ = 1 then tot_al=0;
tot_al + (1 - 1/poids_I) ;
if fini then output ;
run ;

/* merge pour import du paramètre */
data estim_ek ;merge estim_ek(in=un) tot_al(in=deux);
by zus ;
if un and deux;
run ;
```

```

/* Calcul des  $a_i$  */
data estim_ek ;set estim_ek ;
al= (1 - 1/poids_I) / tot_al ;
run ;

/* Calcul du paramètre A */
data GRA ;set estim_ek end=fini;
if _N_= 1 then GRAND_A=0;
GRAND_A + al *ek*poids_I ;
if fini then output ;
keep ZUS GRAND_A;
run ;

data estim_ek ;merge estim_ek(in=un) GRA(in=deux);
by zus ;
if un and deux;
run ;

/* Calcul des  $\sum a_i^2$  */
data SOMMEA2 ; set estim_ek end=fini;
if _N_= 1 then SA2=0;
SA2 + al*al ;
if fini then output ;
keep ZUS SA2;
run ;
data estim_ek ;merge estim_ek(in=un) SOMMEA2(in=deux);
by zus ;
if un and deux;
run ;

/* Calcul de la variance */
data VARIANCE : set estim_ek end=fini;
if _N_= 1 then VARIAN=0;
VARIAN + (1/(1-SA2))*(1-1/poids_I)*((ek*poids_I)-GRAND_A)*((ek*poids_I)-
GRAND_A) ;
if fini then do ; varian = (varian /(NTOT *NTOT)) ;
ecart_type = (varian ) ** 0.5 ;
coeff_variation= ecart_type /0.73 ;

output ;end ;
keep ZUS VARIAN ecart_type coeff_variation ;
run ;

/* Fin */

```

Annexe 9

L'estimateur direct d'Horwitz-Thomson

La population U a pour effectif N .

L'échantillon a pour effectif n .

Dans la suite, nous écrivons \hat{Y} pour l'estimateur de la moyenne des Y .

L'estimateur de la moyenne est :

$$\hat{Y}_h = \frac{1}{N} \sum_{k \in S} \frac{Y_k}{p_k}$$

Qui estime la moyenne dans la population :

$$\bar{Y} = \frac{1}{N} \sum_{k \in U} Y_k$$

Un estimateur de la variance est obtenu par une somme pondérée sur l'échantillon S et non plus sur la population U :

$$\hat{V}(\hat{Y}) = \frac{-1}{2N^2} \left(\sum_{k \in S} \sum_{l \in S} \left(\frac{D_{kl}}{p_{kl}} \right) \left(\frac{y_k}{p_k} - \frac{y_l}{p_l} \right)^2 \right) \quad (1)$$

avec $D_{kl} = p_{kl} - p_k p_l$;

La démonstration de cette formule se trouve par exemple dans la «Théorie des Sondages» de P. Ardilly. Les probabilités du second ordre étant difficiles à calculer (et cette somme représentant n^2 termes à sommer), on cherche à se ramener à une somme simple. Les différentes approximations sont connues.

L'approximation classique de la formule (1), notamment utilisée par Poulpe, donne :

$$\hat{V}(\hat{Y}) = \left[\frac{1}{N^2} \left(\frac{1}{1 - \sum_{l \in S} a_l^2} \right) \left(\sum_{k \in S} (1 - \Pi_k) \left(\left(\frac{y_k}{\Pi_k} \right) - A \right)^2 \right) \right];$$

$$\text{avec} \quad a_l = \frac{(1 - \Pi_l)}{\sum_{k \in S} (1 - \Pi_k)};$$

$$\text{et} \quad A = \sum_{k \in S} a_k \frac{y_k}{\Pi_k}$$

Les développements qui suivent sont issus essentiellement de la «Théorie des sondages» d'Yves Tillé. Le principe de l'approximation repose sur le fait qu'on peut considérer le plan comme à probabilités inégales et à taille fixe, comme un plan de POISSON conditionnel, si ce plan est bien fait, c'est à dire qu'il soit suffisamment aléatoire pour s'approcher du plan à probabilités inégales et de taille fixe qui maximise l'entropie.

Le plan de POISSON :

le Plan de Poisson consiste en un plan à probabilités inégales mais pas de taille fixe. On se donne une variable uniforme u sur l'intervalle $[0,1[$, et on va répéter pour chaque valeur de k variant de 1 à N , N taille de la population les opérations suivantes :

si $u < p_k$, on prend k dans l'échantillon, on passe à $k + 1$.

Ceci permet de construire un plan de POISSON à partir de probabilités d'inclusion p_k qui sont données. La probabilité de tirer un échantillon S de taille quelconque avec cet algorithme sera :

$$p(s) = \prod_{k \in S} p_k \prod_{k \in U-S} (1-p_k)$$

On voit que les unités étant tirées de façon rigoureusement indépendantes les unes des autres, le terme avec $D_{kl} = p_{kl} - p_k p_l$ est nul lorsque $k \neq l$, puisque $p_{kl} = p_k p_l$ et est égal à $p_k - p_k^2$ sinon, ce qui permet de simplifier la formule de la variance pour un tirage à probabilités inégales (de taille quelconque) qui est :

$$V(\hat{Y}) = \frac{1}{N^2} \left(\sum_{k \in U} \sum_{l \in U} \left(\frac{D_{kl}}{p_k p_l} \right) y_k y_l \right) \quad (\text{avec } D_{kl} = p_{kl} - p_k p_l)$$

Cette formule devient , puisque les seuls termes non nuls correspondent à $k = l$,

$$V(\hat{Y}) = \frac{1}{N^2} \left(\sum_{k \in U} \frac{p_k (1-p_k) y_k^2}{p_k^2} \right)$$

qui admet comme estimateur, pour des raisons évidentes :

$$\hat{V}(\hat{Y}) = \frac{1}{N^2} \left(\sum_{k \in S} \frac{(1-p_k) y_k^2}{p_k^2} \right)$$

D'autre part, on peut démontrer que l'entropie :

$$I(p) = - \sum_s p(s) \log p(s), \quad I(p) > 0$$

qui mesure la quantité d'information contenue dans un plan de sondage est Maximale pour un plan de poisson. Pour le prouver, on cherche à déterminer le plan qui maximise l'entropie $I(p)$ sous les contraintes d'un plan de sondage, c'est à dire :

$$\sum_{k \in S, S \subset U} p(s) = p$$

et

$$\sum_{S \subset U} p(s) = 1$$

On écrit l'entropie $I(p)$ comme une fonction (Lagrangien) de $I(p)$, et de paramètres Λ_k et \mathbf{m} en utilisant $\sum p(s) - \Pi_k = 0$ pour tout k de U et $\sum p(s) - 1 = 0$;

$$L(I(p), \Lambda_k, \mathbf{m}) = I(p) + \sum_{k \in U} \Lambda_k \left(\sum_{S \subset U} p(s) - p_k \right) + \mathbf{m} \left(\sum_{S \subset U} p(s) - 1 \right);$$

ou encore :

$$L(I(p), \Lambda_k, \mathbf{m}) = \sum_{S \subset U} p(s) \log(p(s)) + \sum_{k \in U} \Lambda_k \left(\sum_{S \subset U} p(s) - p_k \right) + \mathbf{m} \left(\sum_{S \subset U} p(s) - 1 \right);$$

On peut écrire ce Lagrangien sous la forme :

$$L(I(p), \Lambda_k, \mathbf{m}) = \sum_{S \subset U} \left\{ -p(s) \log(p(s)) + \sum_{k \in U} \Lambda_k p(s) + \mathbf{m} p(s) \right\} - \sum_{k \in U} \Lambda_k \mathbf{p}_k - \mathbf{m} ;$$

En introduisant la fonction :

$$K(I(p), \Lambda_k, \mathbf{m}) = -p(s) \log(p(s)) + \sum_{k \in U} \Lambda_k p(s) + \mathbf{m} p(s) ;$$

On peut écrire la différentielle :

$$d((I(p), \Lambda_k, \mathbf{m})) = \sum_{S \subset U} \frac{\partial K(I(p), \Lambda_k, \mathbf{m})}{\partial p(s)} dp(s) ;$$

Au maximum du Lagrangien, qui est identiquement égal à I(p), sa différentielle est nulle, ce qui est équivalent au système d'équation :

$$\frac{\partial K(I(p), \Lambda_k, \mathbf{m})}{\partial p(s)} = 0, \quad \forall S \subset U$$

c'est-à-dire, en dérivant la fonction :

$$\log p(s) = \sum_{k \in U} \Lambda_k + \mathbf{m} - 1, \quad \forall S \subset U$$

ou encore :

$$p(s) = \exp\left(\sum_{k \in U} \Lambda_k + \mathbf{m} - 1\right) (*)$$

en utilisant les relations :

$$\sum_{k \in S, S \subset U} p(s) = \mathbf{p}$$

et

$$\sum_{S \subset U} p(s) = 1$$

On arrive facilement à déterminer les $\mathbf{p}_k, \Lambda_k, \mathbf{m} - 1$;

Soit

$$\begin{aligned} \mathbf{p}_k &= \exp(\Lambda_k) (1 + \exp(\Lambda_k)) ; \\ \exp(\Lambda_k) &= \mathbf{p}_k (1 - \mathbf{p}_k) ; \\ \exp(1 - \mathbf{m}) &= \prod (1 - \exp(\Lambda_k)) ; \end{aligned}$$

On en déduit le plan de sondage $p(s)$, en utilisant la formule (*) :

$$p(s) = \prod_{k \in S} \mathbf{p}_k \prod_{k \in S} (\mathbf{p}_k - 1)$$

Ce qui constitue le plan de POISSON ; on a donc prouvé que ce plan maximisait l'entropie, c'est à dire que ce plan permet le tirage le plus aléatoire possible ;

Plan à entropie maximum, de taille fixe

Par une démonstration assez semblable, on peut montrer qu'il existe un plan à probabilités inégales, mais de taille fixe, qui a une entropie maximale. Ce plan peut s'écrire, pour n donné :

$$p(s) = \exp\left(\sum_{k \in S} I_k + m(n)\right)$$

Les I_k sont uniques (démonstration effectuée en 1994 par Chen, Dempster et Liu, d'après Tillé). Ces I_k permettent de définir un plan de poisson grâce aux formules :

$$p_k(p) = \frac{\exp(I_k)}{(1 + \exp(I_k))}$$

et qui vérifie :

$$\sum_{k \in S} p_k(p) = n$$

alors que la taille de l'échantillon pour le plan de poisson est ns . On peut écrire le plan qui maximise l'entropie comme étant un plan de poisson conditionnel, la condition portant sur la taille ns .

On aura pour ce plan p , si pp est le plan de POISSON :

$$p(s) = pp(s/ns = n) = \frac{pp(s)}{\sum_{S \subset S_n} pp(s)}$$

$$pp(s) = \prod_{k \in S} p_k(p) (1 - p_k(p))$$

si S_n est l'ensemble des échantillons de taille n .

Les $pp(s)$ s'expriment à partir des $p_k(p)$ et des I_k par

$$pp(s) = \prod_{k \in S} p_k(p) (1 - p_k(p))$$

Il existe un algorithme qui permet, à partir du plan de poisson conditionnel, de passer au plan de taille fixe à entropie maximale, et réciproquement. Il est donc possible, en théorie, de construire un plan de taille fixe idéal. La plus part du temps, on doit s'en écarter pour des raisons purement matérielles.

Estimateur de la variance :

La méthode consiste à calculer la variance d'un estimateur de moyenne, pour un plan de POISSON dont on aurait fixé la taille (plan de POISSON conditionnel) et de considérer que si le plan de taille fixe et à probabilités inégales est bien fait, il doit s'approcher du plan d'entropie maximum, donc du plan de Poisson conditionnel, ce qui nous permettra d'écrire :

si \hat{Y} est l'estimateur direct de moyenne, $Var(\hat{Y}) = \text{variance de poisson } (\hat{Y}|n_s = n)$;

En supposant que (\hat{Y}, n_s) suit une distribution normale, on peut écrire :

$$\text{variance de Poisson } (\hat{Y} | n_s = n) = \text{variance de Poisson } (\hat{Y} + (n_s - n)\mathbf{b})$$

$$\text{avec : } \mathbf{b} = \frac{\text{covariancepoisson}(\hat{Y}, n_s)}{\text{variancepoisson}(\hat{Y} + (n_s - n)\mathbf{b})} ;$$

$$\text{Varpoisson}(n_s) = \sum_{k \in U} \frac{\mathbf{p}_{pk} (1 - \mathbf{p}_{pk}) \mathbf{p}_{pk}^2}{\mathbf{p}_{pk}}$$

puisque $n_s = \sum_{k \in U} \mathbf{p}_{pk}$, \mathbf{p}_{pk} étant la probabilité d'insertion de POISSON.

N étant la taille de la population,

$$\text{covarpoisson}(\hat{Y}, n_s) = \frac{1}{N} \frac{\sum_{k \in U} \mathbf{p}_{pk} (1 - \mathbf{p}_{pk}) y_k}{\mathbf{p}_{pk}} ;$$

En considérant les constantes $b_k = \mathbf{p}_{pk} (1 - \mathbf{p}_{pk})$, on a une bonne approximation par la formule :

$$\text{varapprochée}(\hat{Y}) = \frac{1}{N^2} \sum_{k \in U} \frac{b_k}{\mathbf{p}_k^2} (y_k - y_k^*)^2$$

avec

$$y_k^* = \mathbf{b} \mathbf{p}_k = \frac{\left(\sum_{k \in U} \frac{b_k \mathbf{p}_k}{\mathbf{p}_k} \right)}{\sum_{k \in U} b_k}$$

On peut réécrire cette formule :

$$\text{vap}(\hat{Y}) = \frac{1}{N^2} \left(\left(\frac{\sum_{k \in U} \frac{y_k^2 (b_k - b_k^2)}{\mathbf{p}_k^2}}{\sum_{k \in U} b_k} \right) - \frac{1}{\sum_{l \in U} b_l} \left(\sum_{k \in U} \sum_{l \in U} \frac{y_k y_l b_l}{\mathbf{p}_k \mathbf{p}_l} \right) \right) ;$$

consiste en une somme de termes en nombre égal à l'échantillon, alors que dans la formule générale, on a affaire à n^2 termes.

Or la formule exacte pour la variance de \hat{Y} est :

$$V(\hat{Y}) = \frac{1}{N^2} \sum_{k \in U} \mathbf{p}_k (1 - \mathbf{p}_k) \frac{y_k^2}{\mathbf{p}_k^2} + \frac{1}{N^2} \left(\sum_{k \in U} \sum_{l \in U} \left(\frac{\Delta_{kl}}{\mathbf{p}_k \mathbf{p}_l} \right) \right)$$

On cherche à identifier les deux premiers termes de ces deux sommes, donc à résoudre

$$\left(\frac{b_k - b_k^2}{\sum_{k \in U} b_k} \right) = \mathbf{p}_k (1 - \mathbf{p}_k) ;$$

Il existe différentes méthodes pour résoudre cette équation, comme la méthode du point fixe, nous ne la développerons pas ici. Bref, en transformant cette valeur approchée de la variance, en estimateur, c'est en changeant les sommes sur U en sommes sur S et en remplaçant y^* par \hat{y}^* , on obtient l'estimateur de variance en prenant pour b_k , la valeur :

$$b_k = \frac{(1-p_k)}{\left(1 - \sum_{k \in S} \left[\frac{(1-p_k)}{\sum_{k \in S} (1-p_k)} \right]^2\right)} = \frac{(1-p_k)}{\left(1 - \sum_{l \in S} a_l^2\right)}$$

avec
$$a_l = \frac{(1-p_l)}{\sum_{k \in S} (1-p_k)}$$

en prenant
$$\hat{y}_k^* = p_k \frac{\sum_{l \in S} b_l y_l}{\sum_{l \in S} b_l}$$

on aura :

$$\hat{V}(\hat{Y}) = \frac{1}{N^2} \sum_{k \in S} \frac{b_k}{p_k^2} (y_k - \hat{y}_k^*)^2$$

C'est exactement la formule (1) en remarquant, après un petit calcul que :

$$p_k A = \hat{y}_k^* \quad \text{avec} \quad A = \sum_{k \in S} a_k \frac{y_k}{p_k}$$

