Problèmes théoriques et pratiques de la construction de l'EMEX (*Document de travail*).

Marc CHRISTINE ( ), Laurent WILMS (**)*

(*) INSEE, Unité Méthodes Statistiques

(**) INSEE, RRP

1 Introduction.

Régulièrement, les régions, par la voix des conseils régionaux ou d'autres administrations ou organismes locaux, souhaitent disposer d'extensions régionales[1] pour les enquêtes nationales auprès des ménages réalisées par l'Insee.

Pour les enquêtes les plus récentes, cela a été le cas de l'enquête Logement (tirée en septembre 2001), avec une extension en Nord-Pas de Calais, puis de l'enquête Santé en 2002 (outre celle déjà citée, quatre régions ont bénéficié d'une extension pour cette enquête : Ile-de-France, Provence-Alpes-Côte d'Azur, Champagne-Ardenne, Picardie).

Compte tenu d'une demande croissante pour disposer de résultats significatifs au niveau de la région, il est probable que d'autres enquêtes pourront dans le futur être pourvues d'extensions régionales. Bien entendu, cette affirmation est à nuancer par la nécessité pour les régions concernées de trouver des financements locaux spécifiques pour les extensions. Cette contrainte peut être forte, voire rédhibitoire dans certains cas, mais l'émergence d'une demande régulière en ce sens, même si elle est limitée à quelques régions seulement, légitime la mise en place d'une offre adaptée.

L'objectif de telles extensions est clairement de disposer d'une base de données permettant d'établir des résultats régionaux avec une précision acceptable et un degré de détail permettant de mettre en évidence des effets propres à la région, alors que les échantillons nationaux n'ont pas été construits (sauf exception, type enquête Emploi) dans l'optique de dériver des résultats régionaux[2].

---

[1] Nous rappelons qu'une extension régionale n'a pas pour objectif de répondre à un besoin d'informations locales spécifique à la région, qui est le plus souvent incompatible avec les objectifs de l'enquête nationale. Nous distinguons donc les extensions régionales (d'enquêtes nationales) des « enquêtes purement locales » : pour ces dernières, les thèmes se rapportent directement à une problématique régionale et le champ géographique est souvent de nature infra-régionale ou vise à éclairer des espaces particuliers au sein de la région (une agglomération donnée, les communes voisines d'un axe de transport donné etc...). Naturellement, une extension régionale implique un *questionnaire rigoureusement identique* à celui de l'enquête nationale.

[2] Nous rappelons que l'EM n'a pas, a priori, une représentativité régionale suffisante car le nombre d'Unités primaires tirées par région est trop faible et il est *équilibré à un niveau super-régional* et non régional, sur quatre variables (revenu net imposable et trois tranches d'âge). De plus, il n'y a pas de condition explicite de couverture du territoire de la région par l'échantillon-maître : celui-ci peut impacter la région de manière plus ou moins bien répartie et il peut arriver que certains départements ne contiennent pas d'Unités primaires rurales.

Dans le passé, ces extensions d'enquêtes étaient la plupart du temps gérées directement par les Directions régionales de l'Insee, de manière déconnectée avec l'enquête nationale : échantillon spécifique tiré par la DR, mode de traitement (correction de la non-réponse, redressement, pondération...) autonome, exploitation distincte (risques d'incohérences avec les parties régionales de l'enquête nationale..).

Dans le cadre des travaux de rénovation des échantillons mis en oeuvre après le RP de 1999, l'Unité Méthodes Statistiques a souhaité construire un outil standard permettant une offre cohérente pour les régions, un traitement homogène des demandes d'extension régionale, l'intégration dans les travaux nationaux et une prise en charge globale des questions méthodologiques.

Cet outil, pour être bien adapté aux besoins, nécessite d'abord la construction d'un échantillon ; mais il doit aussi inclure l'organisation des phases de traitement venant en aval : méthodes de correction de la non-réponse, redressement, estimation conjointe sur l'échantillon national et l'extension régionale...

L'étape initiale, à savoir le tirage des échantillons de logements des extensions régionales, va être réalisée grâce à la mise en place d'un outil standard, dit « **E**chantillon **M**aître pour les **Ex**tensions régionales » (**EMEX**).

Cet outil a été mis en place pour la première fois en septembre 2002 pour le tirage des extensions régionales de l'enquête Santé.

Comme son nom l'indique, cet outil procède de la même philosophie d'ensemble que celle de l'échantillon-maître (EM) : tirage à deux degrés, concentration de la localisation des échantillons pour limiter les coûts de déplacement, souci d'une bonne «représentativité» régionale EM + EMEX, garante de la qualité des résultats..., tout en assurant un enrichissement de l'échantillon-maître national grâce à une exploitation conjointe.

La mise en place de l'EMEX a révélé quelques difficultés théoriques, notamment pour le premier degré de tirage. Il s'agissait en effet de tirer un échantillon d'unités primaires «régionales» astreintes à certaines conditions (stratification ou équilibrage), *conditionnellement* au tirage des unités primaires « nationales » de l'EM, qui avait été fait une fois pour toutes en 2000.

Le papier exposera les questions théoriques et les différentes voies pour y répondre, notamment comment concilier les contraintes d'équilibrage et le respect de probabilités d'inclusion finales données, dans le cadre de ce tirage conditionnel. Il décrira les solutions retenues in fine et la manière dont il sera envisagé de traiter conjointement les deux échantillons, à la fois pour les exploitations nationales et celles des extensions.

Dans le paragraphe 2, nous donnons les principes généraux de constitution de l'EMEX. Puis, dans le 3, nous abordons la question théorique de l'équilibrage pour le tirage des Unités primaires de l'EMEX et les questions d'inférence régionale à partir de ces deux échantillons-maîtres.

La partie 4 est consacrée aux tirages à l'intérieur des Unités primaires, la partie 5 à un point particulier qui est le traitement de la construction neuve dans ce cadre. Dans la partie 6, quelques éléments sont donnés sur les développements informatiques.

La partie 7, enfin, fournit des indications sur les traitements qui seront opérés en aval.

2 Principes généraux de constitution de l'EMEX.

Le tirage des extensions mis en place se déroule selon un mode proche de celui du tirage des logements dans l'échantillon-maître[3], c'est-à-dire :

- un premier degré consistant en un tirage d'unités primaires[4] spécifiques (dites UP-EMEX), en nombre fixé à l'intérieur de chaque strate définie par le croisement de la région et de la tranche de taille d'unité urbaine (selon des modalités regroupées[5])
- et un second degré dans lequel, pour une enquête à extension, les logements sont tirés dans les UP-EMEX des régions bénéficiant d'une extension[6].

Les UP-EMEX tirées doivent donc alimenter ainsi l'ensemble des extensions régionales pour les enquêtes nationales concernées. Elles permettent, par construction, de conserver une *certaine concentration géographique des enquêtes* par rapport à une solution dans laquelle l'échantillon d'une extension serait dispersé dans toute la région. Cette concentration, d'une part, est compatible avec l'organisation traditionnelle du réseau d'enquêteurs, d'autre part, permet de contenir les coûts de collecte.

Par ailleurs, fixer ex-ante les unités primaires peut faciliter le suivi des logements neufs, donc assurer la possibilité d'enquêter aussi ces logements. *Cette possibilité est toutefois tributaire des moyens que les DR susceptibles d'accueillir des extensions peuvent y consacrer et a nécessité la mise en place d'un système un peu différent du système national (cf. § 6) .*

Les grandes lignes adoptées pour l'EMEX ont été énoncées et validées sous forme de principes.

---

[3] Cf. VII èmes JMS, 4-5 décembre 2000, « Echantillons Maître et Emploi », G. BOURDALLE, M. CHRISTINE, L. WILMS, Insee-Méthodes n° 100 (tome 1), pages 139-241.

[4] La définition des unités primaires de l'EMEX est la même que celle utilisée pour l'échantillon-maître.

[5] On utilise le concept de « *strate de gestion* ». Chaque strate de gestion est caractéristique d'un degré d'urbanisation. La strate de gestion 0 contient les UP dites rurales construites sur la base de regroupements de communes rurales contiguës (ce qui correspond le plus souvent à des cantons ruraux). Les UP des strates de gestion 1, 2 et 3 sont des unités urbaines (UU) classées dans ces strates selon leur population :

Strate 1 : UU de moins de 20 000 habitants

Strate 2 : UU de 20 000 à 100 000 habitants

Strate 3 : UU de plus de 100 000 habitants (hors UU de Paris)

Strate 4 : UU de Paris.

[6] In fine, on dispose donc, dans une région bénéficiant d'une extension, de l'échantillon de logements issu des UP-EMEX et de celui issu des UP-EM.

2.1 Les règles générales de constitution de l'EMEX.

<u>Principe 1</u> : règles générales des tirages d'extensions régionales.

- le tirage d'une extension régionale n'a pas vocation à prendre en compte une demande locale d'informations spécifique à la région et à elle seule.

- les extensions régionales d'enquêtes nationales sont tirées hors échantillon-maître.

- les tirages s'effectuent selon un mode analogue à celui de l'échantillon-maître national : tirage d'unités primaires spécifiques, de districts (seulement en strates de gestion 2, 3 et 4) puis de logements. Ainsi, est constitué un « échantillon-maître d'extension », EMEX, dans lequel seront tirés les logements relatifs à ces enquêtes.

- les unités primaires tirées pour l'EMEX sont fixées une fois pour toutes pour l'ensemble des extensions régionales.

<u>Principe 2</u> : tirage des unités primaires EMEX selon les strates de gestion.

En strates de gestion 0, 1 et 2.

Au sein d'une région, les unités primaires de l'EMEX sont tirées obligatoirement dans la base des unités primaires qui ont été constituées lors de la construction de l'échantillon-maître national, d'où l'on a exclu les unités primaires tirées pour ce dernier[7] (il y a donc disjonction totale de l'EMEX avec l'EM).

En strates de gestion 3 et 4.

Les unités primaires sont des Unités urbaines. Elles sont toutes retenues de manière exhaustive, comme elles l'ont été pour l'échantillon-maître national.

<u>Principe 3</u> : une gestion intégrée de l'organisation des travaux.

- la constitution de la base de logements EMEX (c'est-à-dire des Unités primaires qui le composent) relève de la responsabilité de l'UMS et non des Directions régionales de l'INSEE.

- l'équipe informatique en charge de l'EM a développé un module dans la chaîne de tirage permettant, pour une région à extension, de tirer conjointement les logements dans la base constituée par l'EM et l'EMEX.

[7] En effet, les Unités primaires de l'EM national pourraient s'épuiser prématurément si les extensions y étaient tirées.

2.2 Dimensionnement de l'EMEX.

Les discussions avec les responsables des futures enquêtes à extensions connues au moment du lancement de la réflexion ont conduit à chercher à anticiper le dimensionnement de l'EMEX. Il s'est dégagé que la taille envisagée de l'échantillon de l'extension dans une région donnée, en nombre de logements, serait en général proche de celle de l'échantillon régional hors extension[8] (c'est-à-dire pour la partie tirée dans l'échantillon-maître national) : en d'autres termes, une région à extension voit sa taille d'échantillon doubler par rapport à une situation où elle ne bénéficierait pas d'une extension (taux de sondage moyen double)..

Par ailleurs, des paramètres standards utilisés pour l'EM restent valides :

- le taux de sondage d'une enquête standard sans extension est de 1/2000
- la charge moyenne par enquêteur est de 20 FA

Ces hypothèses ayant été utilisées pour la constitution de l'EM, on peut comprendre que la prise en compte du premier principe conduise à définir une base EMEX de taille équivalente à celle de l'EM. Le taux d'épuisement de la base EMEX sera, de toute évidence, plus faible que celui de l'EM et, au pire, égal dans le cas où une région solliciterait une extension à chacune des enquêtes nationales, ce qui est peu réaliste.

A partir de ces données, on va définir combien on doit sélectionner d'Unités primaires pour l'EMEX.

2.2.1 En strates de gestion 0 et 1.

Il s'agit des unités primaires rurales (strate 0) et des unités urbaines de moins de 20.000 habitants (strate 1).

La taille de l'échantillon d'unités primaires de l'EMEX doit être justifiée par la taille des extensions. On a envisagé en fait deux hypothèses :

- soit doubler la taille totale (EM + EMEX) de l'échantillon d'UP (par rapport au nombre d'UP de l'EM)
- soit la tripler.

On peut penser que la majorité des extensions régionales conduiront à un doublement de la taille de l'échantillon de logements issu du tirage national : un doublement du nombre d'UP serait alors bien adapté pour que chaque enquêteur conserve le même nombre moyen de fiches adresses (soit une vingtaine). Mais, pour se ménager des moyens de manœuvre, le triplement de la taille de l'échantillon d'UP donnerait tout le confort souhaité (si l'extension régionale allait jusqu'à tripler par exemple).

Il a finalement été décidé de tripler la taille des échantillons d'UP..

Le choix d'un triplement du nombre d'UP-EM et UP-EMEX, sur les strates de gestion 0 et 1, se justifie pour trois raisons :

[8] Cela dépend en général du budget total alloué à l'extension, mais un calibrage ex-ante de l'EMEX, fixé à l'équivalent de l'EM, semble refléter les pratiques moyennes en termes d'extension.

1) un argument de précision : plus on augmente le nombre d'UP tirées, plus on augmente la précision des estimateurs. Il est important de faire porter cet effort en priorité sur les strates de gestion 0 et 1 où les UP sont de grandes tailles et peu nombreuses. De plus, dans le cas d'une extension régionale de grande taille (par exemple, taux de sondage triple du taux de sondage des régions sans extension), il est légitime de chercher à augmenter corrélativement le nombre d'UP à impacter dans la région à extension.

2) un argument de charge par enquêteur dans le cas d'une extension régionale de grande taille : si le nombre d'UP-EM+UP-EMEX est trop faible, on pourrait être dans la situation où la charge par UP est trop importante pour un enquêteur et trop faible si l'on recrute un second enquêteur.

3) un nombre d'UP EM+EMEX trop important vis-à-vis d'une extension régionale de petite taille n'est pas gênant et n'implique pas une charge de FA par UP trop faible : en effet, la chaîne de tirage d'enquêtes propose une option permettant, dans ce cas, de ne pas puiser les logements dans l'ensemble des unités primaires sélectionnées au 1er degré mais de réaliser un sous-échantillonnage, par sondage aléatoire simple, des UP (EM et EMEX) qui seront effectivement sollicitées. Ceci permet de garantir une charge minimale de FA par UP. Cette option fait cependant perdre d'éventuelles propriétés d'équilibrage lors du tirage des UP.

2.2.2 En strate de gestion 2.

Il s'agit des Unités Urbaines (UU) de 20 000 à 100 000 habitants.

Données : Il y a 180 unités urbaines de strate 2 en France. 93 sont actuellement dans l'échantillon-maître.

L'annexe 1 fournit la liste des 87 unités qui ne sont PAS dans l'échantillon-maître (tri par région).

Compte tenu du principe qu'une extension régionale conduira en moyenne à un doublement de l'échantillon de logements (par rapport à l'effectif régional hors extension), il n'est guère d'autre possibilité que prendre pour l'EMEX les UP qui n'ont pas été tirées pour l'EM.

Pour cette strate, il a donc été décidé de retenir pour l'EMEX toutes les UP non tirées dans l'EM, c'est-à-dire de doubler approximativement le nombre d'UP sélectionnées pour l'EM.

Ce choix permet en outre d'obtenir une meilleure précision sur les extensions régionales ; de surcroît, l'EMEX n'a pas vocation à fournir des estimations par département, mais il est souhaitable d'avoir une bonne couverture géographique de la région. S'ajoute une communication plus facile, car il peut être gênant, pour un maître d'ouvrage, que des villes importantes, par leur taille ou leur statut (une préfecture par exemple), ne participent pas à une enquête régionale.

Néanmoins, ce choix entraîne deux conséquences :

- Evidemment un coût plus élevé, soit en déplacement, soit en recrutement d'enquêteurs.

- Des effectifs tirés par agglomération désormais proportionnels à la taille de l'agglomération (donc *variables* d'une UU à l'autre). De ce fait, si l'extension est de taille modeste, cela conduira à un faible nombre de FA pour les agglomérations les plus petites (taux régional global de 1/1500 dans une agglomération de 15.000 logements : 10 fiches-adresses).
- Une solution inverse (mais non retenue) aurait consisté à procéder à un échantillonnage des UU (dans le cas des extensions de petite taille), ce qui aurait entraîné une charge d'enquête *constante* par UU, au moins pour les plus petites (de l'ordre d'une vingtaine de FA par enquêteur pour une enquête de taille « moyenne »). Théoriquement, on peut imaginer des traitements différents selon les DR, mais la justification n'est pas évidente alors que cela entraînerait par ailleurs de sérieuses complications de programmation.

On peut résumer ces discussions sur le dimensionnement sous la forme du principe suivant :

<u>Principe 4</u> : dimensionnement de l'échantillon d'UP.

En strates de gestion 0 et 1.

Au sein de chaque région, le nombre d'UP-EMEX dans chaque strate de gestion est égale au double du nombre d'UP-EM de la strate.

En strate de gestion 2.

Au sein d'une région concernée par une extension, les unités primaires de l'extension sont sélectionnées de manière exhaustive dans le complémentaire de l'ensemble des unités primaires qui ont été tirées pour l'échantillon-maître national, ce qui constitue un doublement du nombre d'UP disponibles.

3 Le tirage des Unités primaires de l'EMEX : la question de l'équilibrage.

La constitution de l'EM, après la phase de construction des UP, avait nécessité de procéder à un tirage aléatoire de celles-ci. Ce tirage des UP constituait le 1er degré de tirage de l'échantillon-maître. Dans les strates de gestion 3 et 4, toutefois (unités urbaines de plus de 100.000 habitants), l'intégralité des UP constituées avaient été retenues.

Ce tirage s'était fait selon les quatre principes suivants :

- stratification par région et « strate de gestion ».
- nombre d'UP fixé à l'intérieur de chacune des strates.

- probabilité de tirage des UP proportionnelle à leur taille exprimée en nombre de logements principaux au RP 1999.
- équilibrage sur différentes variables : tranches d'âge, revenu, nombre de logements...

Ce tirage avait été réalisé à l'aide de la Macro CUBE.

On rappelle que l'équilibrage consiste à imposer qu'un estimateur de type HORVITZ-THOMSON (H-T) du total de certaines variables d'intérêt, fabriqué à partir des unités tirées, prenne une valeur identique à celle (connue) du total sur l'ensemble de la population. L'objectif est de faire en sorte que l'échantillon tiré constitue un meilleur modèle réduit de la population de référence au vu des variables d'intérêt mises en oeuvre ; par abus de langage, on dira que ce tirage assure une meilleure « *représentativité* ».

La question se pose dans les mêmes termes pour le tirage des UP-EMEX mais conduit à une difficulté théorique importante qui va être explicitée dans le présent paragraphe.

On notera tout d'abord que la strate de gestion 2 devenant exhaustive en termes d'UU (comme le sont les strates 3 et 4), après tirage de l'EMEX, il n'y a pas de problème d'équilibrage au niveau du tirage des UP dans ces strates.

Les questions d'équilibrage se trouvent reportées au niveau du tirage des districts dans chacune de ces UU (cf. § 4).

La question traitée dans cette partie ne se pose donc que dans les strates de gestion 0 et 1 : ***peut-on tirer les Unités primaires de l'EMEX tout en assurant des conditions d'équilibrage pour l'ensemble EM + EMEX au niveau de la région ?***

On notera tout d'abord que ce problème eût été plus simple à résoudre si l'on avait tiré simultanément les UP de l'EM et de l'EMEX (cf. annexe 2). L'EMEX étant une extension nouvelle au projet Echantillon-maître, sa prise en compte n'avait pas été prévue au démarrage.

Pour mettre en oeuvre des conditions d'équilibrage, il convient d'examiner comment construire des estimateurs sans biais de totaux à partir des deux échantillons : EM et EMEX. C'est la question de *l'inférence régionale.*

3.1 Eléments de réflexion sur l'inférence régionale à partir des Unités primaires EMEX et EM.

Disposant de deux échantillons relatifs à la région, l'un correspondant à la partie régionale de l'EM, l'autre issue de l'EMEX, comment combiner les observations disponibles pour estimer sans biais un total régional ? Plus concrètement, comment pondèrera-t-on les unités primaires de l'ensemble EM + EMEX pour estimer sans biais un total ?

A priori, il paraîtrait absurde de ne pas utiliser les données issues de l'un et l'autre de ces échantillons pour estimer un total régional. Par ailleurs, il peut aussi sembler souhaitable d'utiliser les observations issues de l'extension pour améliorer la précision d'un estimateur national.

En effet, outre qu'il peut paraître plus avantageux d'utiliser un volume plus important de données, une solution différente de cette dernière conduirait à affecter des poids différents aux observations issues de l'EM selon qu'elles sont utilisées pour une estimation nationale (hors extension régionale) ou régionale. Ceci pourrait poser quelques difficultés pratiques d'utilisation des fichiers (avec plusieurs jeux de pondération).

Trois questions principales se posent en matière d'inférence « régionale » :

- quelles conditions d'équilibrage imposer ?

- comment construire des estimateurs et lesquels choisir ?
- comment tirer les Unités primaires de l'EMEX ?

La particularité fondamentale du contexte de l'EMEX est qu'il faut tirer des UP *conditionnellement à la réalisation du tirage de l'EM.* Pour réaliser ce tirage, il va donc falloir se donner une loi conditionnelle de tirage. A priori, on doit avoir une maîtrise au moins partielle de cette loi. Elle sera résumée (avec perte d'information) par les *probabilités d'inclusion conditionnelles* d'ordre 1 des Unités primaires tirées pour l'EMEX.

Dans ce cadre, *on considère comme donnée intangible la loi de tirage des Unités primaires de l'EM* ; celle-ci n'est connue également que partiellement par l'intermédiaire des probabilités d'inclusion d'ordre 1 de ces Unités primaires.

Lorsque l'on dispose de ces deux lois, on peut bâtir différents estimateurs sans biais pour un total régional. La forme générale choisie sera :

$$\hat{T} = \sum_{i \in S_1} a_i(S_1) Y_i + \sum_{j \in S_2} b_j(S_1) Y_j$$

L'annexe 3 présente le détail de cette approche que nous baptisons «***d'échantillonnages successifs »*** et qui est adaptée au cas où les deux échantillons d'unités primaires sont disjoints.

Deux familles principales d'estimateurs apparaissent alors, à partir de la forme générale ci-dessus :

- les estimateurs à <u>coefficients fixes</u>, c'est-à-dire dans lesquels les pondérations des Unités sont déterminées ex-ante pour toute unité de la population et s'appliquent à elles à partir du moment où elles sont sélectionnées. C'est l'approche classique des estimateurs de Horvitz-Thomson.
    - o estimateur de Horvitz-Thomson relatif à la partie régionale de l'EM (sans intérêt).
    - o estimateur de Horvitz-Thomson relatif au seul EMEX, en utilisant les probabilités d'inclusion *finales* des unités tirées dans l'EMEX.
    - o <u>combinaison barycentrique</u> des deux précédents : $\hat{T} = a \sum_{i \in S_1} \frac{Y_i}{?^1_i} + (1-a) \sum_{i \in S_2} \frac{Y_i}{?^2_i}$
    - o <u>estimateur de Horvitz-Thomson utilisant la réunion de l'EM et de l'EMEX</u> : : $\hat{T} = \sum_{i \in S_1 \cup S_2} \frac{Y_i}{?_i}$

      Ceci nécessite de pouvoir calculer les probabilités d'inclusion finales Π_i d'une Unité primaire dans l'échantillon global[9].

[9] Le calcul de ces estimateurs nécessite en effet de connaître explicitement la loi de tirage de l'EM afin de pouvoir calculer les probabilités d'inclusion finales d'une Unité primaire dans l'EMEX ou dans la réunion de l'EM et de l'EMEX à partir de la loi conditionnelle de tirage de l'EMEX. Sauf dans le cas de tirages simples où la loi est complètement déterminée, ce calcul n'est en général pas possible.

- une approche plus approfondie, dans laquelle, les pondérations des unités dans les deux échantillons sont aléatoires. En se restreignant à la configuration dans laquelle des pondérations sont astreintes à ne dépendre que des réalisations du 1er échantillon et non du 2ème), on obtient une famille générique d'estimateurs de la forme :

$$\hat{T} = \sum_{i \in S_1} \frac{a_i Y_i}{\Pi_i^1} + \sum_{i \in S_2} \frac{(1 - a_i) Y_i}{\Pi_i^{2/S_1} \left(1 - \Pi_i^1\right)}$$

Dans cette formule, a_i est un élément de [0,1], Π_i^1 représente la probabilité d'inclusion d'ordre 1 de l'unité i dans le 1er échantillon (EM) et Π_i^{2/S_1} est la variable aléatoire dont les réalisations sont les probabilités d'inclusion d'ordre 1 de l'unité i dans le 2ème échantillon (EMEX), *conditionnellement à la réalisation* s_1 *de l'échantillon* S_1.

Ces estimateurs ont l'avantage d'être calculables explicitement à partir de la connaissance de ces deux classes de probabilités d'inclusion.

A partir de cette forme générique, on peut dériver deux estimateurs :

- l'un est un estimateur de type Horvitz-Thomson issu de la loi conditionnelle de tirage de l'EMEX, de la forme : $\hat{T}_1 = \sum_{i \in S_1} Y_i + \sum_{i \in S2} \frac{Y_i}{\Pi_i^{2/S_1}}$
- l'autre a pour expression :

$$\hat{T}_2 = a \sum_{i \in S_1} \frac{Y_i}{\Pi_i^1} + (1 - a) \sum_{i \in S_2} \frac{Y_i}{\Pi_i^{2/S_1} \left(1 - \Pi_i^1\right)}$$

Bien entendu, des combinaisons barycentriques de ces estimateurs peuvent aussi être proposés. Les coefficients assurant l'amélioration par rapport aux estimateurs initiaux nécessitent néanmoins de connaître la variance de ces derniers.

Pour écrire concrètement des conditions d'équilibrage relatives au tirage des UP, on utilisera l'un ou l'autre des estimateurs identifiés ci-dessus.

Remarque :

Dans l'approche théorique de ce problème, Y_i est une variable d'intérêt définie sur l'Unité primaire i, supposée connue.

- Lorsque l'on écrit des conditions d'équilibrage relatives au 1er degré de tirage, la ou les variables Y_i seront en pratique des totaux connus de variables d'intérêt définies sur les unités secondaires (logements) de l'unité primaire i.
- Si l'on voulait écrire la forme complète de l'estimateur d'un total incluant les différents degrés de tirage, il conviendrait de remplacer les Y_i par un estimateur relatif au 2ème degré de tirage au sein de l'Unité primaire i.

3.2 Choix des probabilités d'inclusion.

La question suivante qui se pose est celle de la détermination des probabilités d'inclusion (d'ordre 1). On a vu qu'il y a plusieurs concepts utilisables :

- probabilités d'inclusion dans le 1er échantillon (EM) : ces probabilités sont données une fois pour toutes, proportionnelles à la taille en termes de nombre de logements principaux au RP de 1999 : Π_i^1.
- probabilités d'inclusion dans le 2ème échantillon (EMEX) conditionnellement à la réalisation du premier : Π_i^{2/S_1}.
- probabilités finales d'inclusion dans le 2ème échantillon : Π_i^2.
- probabilités finales d'inclusion dans la réunion des deux échantillons : Π_i.

Ces différentes probabilités sont reliées entre elles par des relations démontrées dans l'annexe 3 :

$\boxed{\Pi_i^2 = E\Pi_i^{2/S_1}}$

$\boxed{\Pi_i = \Pi_i^1 + E\,(\Pi_i^{2/S_1} 1_{i \notin S_1})}$

$\boxed{\Pi_i^1 + \Pi_i^2 = \Pi_i}$, dans le cas d'échantillons disjoints.

Sur quels paramètres a-t-on des degrés de liberté ? On se heurte ici à une difficulté fondamentale :

- il peut être plus intéressant de s'imposer des contraintes sur les probabilités d'inclusion finales. En effet, ces conditions se déduiront du mode de tirage aux degrés ultérieurs et des conditions que l'on souhaitera imposer pour la pondération des unités finales (logements) : autopondération (équiprobabilité) conditionnelle ou finale ? autopondération réalisée au sein de l'EMEX mais avec une valeur différente de celle des logements issus de l'EM ?
- mais le tirage sur lequel on a prise est le tirage du 2ème échantillon conditionnellement à la réalisation du tirage du 1er. Ce tirage conditionnel impose de connaître ou de se donner les probabilités d'inclusion conditionnelles.

L'autre difficulté subséquente est que l'on voit, d'après les formules ci-dessus, qu'il n'y a pas de relation simple entre ces probabilités, marginales et conditionnelles. La connaissance des probabilités d'inclusion conditionnelles, pour passer aux probabilités d'inclusion finales, nécessite de connaître la loi complète de tirage du 1er échantillon et pas seulement les probabilités d'inclusion d'ordre 1. Inversement, pour des probabilités finales données, il n'est pas possible de déterminer en général toutes les probabilités conditionnelles qui conduisent à ces probabilités finales.

Il existe cependant un cas simple[10] où l'on peut déterminer des solutions explicites pour ce problème. Il s'agit du cas où l'on suppose que $?_i^{2/s_1}$ ne dépend de s_1 que par l'intermédiaire de l'indicatrice $\mathbf{1}_{i \in s_1}$, c'est-à-dire : $?_i^{2/s_1} = ?_i$ si i appartient à s_1

μ_i si i n'appartient pas à s_1.

L'annexe 3 détaille le mode d'obtention de la solution :

$$\boxed{\Pi_i = (1 - \mu_i)\ \Pi_i^1 + \mu_i}$$

Ainsi, par exemple, dans le cas où l'on souhaiterait que les probabilités *finales* d'inclusion des unités primaires dans la réunion des deux échantillons soient, comme elles le sont pour le seul EM, proportionnelles à leur taille, avec un effectif déchantillon fixé (à l'intérieur d'une région donnée), cela imposerait la condition :

$$\forall\ i \in \boldsymbol{P} : \Pi_i = k\frac{T_i}{T}$$

où T est la taille de l'unité i, T la taille totale de la population (somme des tailles des unités) et k le nombre d'unités tirées.

Avec la condition $\Pi_i = k\frac{T_i}{T}$ et une condition analogue pour l'EM : $\Pi_i^1 = a\frac{T_i}{T}$, où a est le nombre d'unités tirées dans l'EM, on obtient :

$$\boxed{\mu_i = \frac{\Pi_i - \Pi_i^1}{1 - \Pi_i^1} = \frac{(k-a)T_i}{T - aT_i}}$$

On obtient bien des probabilités admissibles, à valeurs dans [0,1].

[10] Proposé par P. ARDILLY.

3.3 Deux contraintes inconciliables : équilibrage et respect de probabilités d'inclusion données.

Les développements précédents montrent que toutes les formes d'estimateurs que l'on peut retenir et, par suite, toutes les équations d'équilibrage que l'on souhaite introduire pourront s'écrire sous une forme générique.

On retiendra la formulation :

Estimateur d'un total : $\hat{T} = \sum_{i \in S_1} a_i(S_1) Y_i + \sum_{j \in S_2} b_j(S_1) Y_j$

Une équation d'équilibrage s'écrira alors sous la forme : $\hat{T} = T$, où T est le vrai total connu sur la population au sein de laquelle on tire l'échantillon.

Dans le cadre des échantillonnages successifs, on est confronté à un problème particulier et tout à fait pénalisant. En effet, le tirage des unités primaires de l'EMEX se fait conditionnellement au tirage de celles de l'EM. La ou les équations de calage doivent faire intervenir l'ensemble des unités tirées dans les deux échantillons.

Il s'agira donc de tirer un échantillon S_2 , selon une loi de tirage qui sera résumée par les probabilités d'inclusion conditionnelles d'ordre 1, Π_i^{2/S_1} , et astreint à des conditions d'équilibrage de la forme :

$$\sum_{j \in S_2} b_j(S_1) Y_j \;=\; T - \sum_{i \in S_1} a_i(S_1) Y_i \;,$$

où les familles de coefficients sont définies fonctionnellement à partir des deux lois de tirage (la loi de tirage de l'EM et la loi conditionnelle de tirage de l'EMEX) de telle sorte que l'estimateur fondé sur la réunion des deux échantillons soit un estimateur sans biais, linéaire en les observations, *pour tout total de variables d'intérêt*.

La méthode du CUBE qui permet de réaliser des tirages équilibrés traite de conditions d'équilibrage de la forme : $\sum_{i \in s} \frac{x_i}{\Pi_i} = \sum_{i \in P} x_i$, où s est un échantillon générique tiré dans une population ***P***, c'est-à-dire que *le total sur lequel on équilibre doit être le total des observations, sur la population, de la variable (éventuellement vectorielle) d'équilibrage.*

La forme de l'estimateur utilisé (ici HORVITZ-THOMSON) ne dépend pas fonctionnellement des variables d'équilibrage : elle est calculée de telle sorte que l'estimateur obtenu soit un *estimateur sans biais du total de n'importe quelle variable Y.* On est alors dans la configuration où : $E\left[\sum_{i \in s} \frac{Y_i}{\Pi_i}\right] = \sum_{i \in P} Y_i$, l'espérance étant prise par rapport à la loi de tirage de l'échantillon.

Transposé au cas présent, cela voudrait dire que l'on ne pourrait traiter que des conditions d'équilibrage de la forme : $\sum_{i \in S_2} \frac{x_i}{\Pi_i^{2/S_1}} = \sum_{i \in U} x_i$,

où U est la population dans laquelle est tiré l'échantillon S_2 (concrètement : ***P*** - S_1) et avec la condition :

E [$\sum_{i \in S_2} \frac{Y_i}{\Pi_i^{2/S_1}}$ / S_1] = $\sum_{i \in U} Y_i$, pour toute variable Y dont on estimerait le total à partir du tirage conditionnel du 2ème échantillon (l'espérance est ici prise *vis-à-vis de la loi conditionnelle de* S_2 *sachant* S_1).

Clairement, et sauf cas très particuliers, le problème posé n'entre pas dans les conditions d'application du tirage CUBE.

On ne maîtrise pas à ce jour de méthode permettant à la fois d'équilibrer et de respecter des probabilités d'inclusion (finales ou conditionnelles) données. Ceci va conduire à rechercher des solutions approchées.

3.4 Une solution approchée : « l'équilibrage inverse[11] ».

On va se placer dans le cadre où l'on utilise un estimateur de Horvitz-Thomson sur la réunion des deux échantillons, avec une probabilité finale d'inclusion Π_i donnée.

L'équation d'équilibrage sur une variable X devrait s'écrire : $\sum_{i \in S_1} \frac{x_i}{\Pi_i} + \sum_{i \in S_2} \frac{x_i}{\Pi_i} = \sum_{i \in P} x_i$ (1).

Comme elle n'est pas assurée a priori, on va chercher de nouvelles probabilités d'inclusion finales, notées $\tilde{\Pi}_i$, proches des Π_i au sens d'une certaine distance, concrètement en cherchant à minimiser un critère de la forme : $\sum_{i \in P} d(\tilde{\Pi}_i, \Pi_i)$, tout en respectant l'équation d'équilibrage (1), où ***P*** est l'ensemble des UP constituant une strate de gestion (0 ou 1) d'une région donnée et où d désigne une distance (distance euclidienne ou du χ^2 en pratique).

Mais l'équation d'équilibrage doit s'intégrer dans le tirage conditionnel de S_2 sachant S_1 (car les UP-EM ont déjà été tirées).

On cherche donc une forme équivalente à l'équation (1), qui s'interprète bien comme une condition d'équilibrage à respecter lors du tirage conditionnel de S_2 sachant S_1.

Pour cela, on va écrire l'équation (1) sous la forme :

$$\sum_{i \in S_2} \frac{z_i}{\tilde{\Pi}_i^{2/S_1}} = \sum_{i \in P} x_i - \sum_{i \in S_1} \frac{x_i}{\tilde{\Pi}_i}$$

[11] Dans la procédure classique d'équilibrage, on peut tirer n'importe quel échantillon avec des probabilités d'inclusion ***fixées ex-ante*** en astreignant l'échantillon à vérifier des contraintes d'équilibrage. Dans la présente procédure, on procède de manière « inverse » : on détermine des probabilités d'inclusion qui assurent l'équation d'équilibrage et ***dépendent, de ce fait, des variables d'équilibrage choisies*** et des valeurs qu'elles prennent dans la population.

où l'on a posé : $z_i = x_i \frac{\tilde{\Pi}_i^{2/S_1}}{\tilde{\Pi}_i}$ et où $\tilde{\Pi}_i^{2/S_1}$ désigne une nouvelle probabilité conditionnelle de tirage de l'UP n° i sachant S_1.

$\rightarrow$ Pour que cette équation s'interprète comme une véritable contrainte d'équilibrage sur le total des variables z_i sur la population, au sein de laquelle on réalise ce tirage conditionnel (soit ***P*** - S_1), il faut qu'elle s'écrive sous la forme :

$$\sum_{i \in S_2} \frac{z_i}{\tilde{\Pi}_i^{2/S_1}} = \sum_{i \in P - S_1} z_i \ .$$

Ces conditions sont satisfaites si et seulement si :

$$\sum_{i \in P} x_i \ - \sum_{i \in S_1} \frac{x_i}{\tilde{\Pi}_i} = \sum_{i \in P - S_1} z_i = \sum_{i \in P - S_1} x_i \frac{\tilde{\Pi}_i^{2/S_1}}{\tilde{\Pi}_i} \ .$$

Si l'on prend le cas simple décrit au § 3.2 où l'on a :

$?_i^{2/s_1} = ?_i$ si i appartient à s_1

μ_i si i n'appartient pas à s_1,

on aura la relation, pour tout i n'appartenant pas à S_1 :

$$\tilde{\Pi}_i^{2/S_1} = \frac{\tilde{\Pi}_i - \Pi_i^1}{1 - \Pi_i^1} \ .$$

L'équation de contrainte s'écrira donc exclusivement en fonction des $\tilde{\Pi}_i$:

$$\sum_{i \in P} x_i \ - \sum_{i \in S_1} \frac{x_i}{\tilde{\Pi}_i} = \sum_{i \in P - S_1} x_i \frac{1 - \frac{\Pi_i^1}{\tilde{\Pi}_i}}{1 - \Pi_i^1} \ .$$

Finalement, dans ce cas, le problème revient à résoudre le programme (P) de recherche des $\tilde{\Pi}_i$:

Min $\sum_{i \in P} d(\tilde{\Pi}_i, \Pi_i)$

sous la contrainte :

$$\sum_{i \in P} x_i \ - \sum_{i \in S_1} \frac{x_i}{\tilde{\Pi}_i} = \sum_{i \in P - S_1} x_i \frac{1 - \frac{\Pi_i^1}{\tilde{\Pi}_i}}{1 - \Pi_i^1} \ .$$

Au total, la résolution de ce programme permet de déterminer des probabilités d'inclusion finales (et, par suite, des probabilités d'inclusion conditionnelles) telles que la condition d'équilibrage imposée sur la réunion des deux échantillons s'interprète effectivement comme une condition d'équilibrage sur une nouvelle variable. Celle-ci est alors automatiquement assurée dès que l'on tire le 2ème échantillon selon ces probabilités d'inclusion conditionnelles. ***D'où le nom « d'équilibrage inverse » que l'on peut attribuer à la procédure.***

En pratique, on pourra équilibrer sur les 4 variables utilisées lors du tirage de l'EM (revenu fiscal et 3 tranches d'âge), d'où en réalité 4 contraintes du type de celles écrites ci-dessus. On ajoutera également la contrainte de taille fixe donnée par :

$$\sum_{i \in P} \tilde{\tilde{\Pi}}_i = k.$$

Cette approche pose toutefois la question de l'existence de solutions dans admissibles (avec tous les $\tilde{\tilde{\Pi}}_i$ dans $]\Pi_i^1, 1[$ et peut présenter des difficultés de résolution de programmes mathématiques complexes.

3.5 Le tirage effectif des UP de l'EMEX.

Finalement, on a été confronté à l'alternative suivante :

- soit faire un tirage équilibré selon la méthode de l'équilibrage inverse mais sans respecter exactement des probabilités d'inclusion fixées (donc en acceptant une modification des probabilités de sélection des unités).
- soit renoncer à l'équilibrage et procéder à un tirage systématique classique.

Ce qui suit expose les tests de tirages des UP en strate de gestion 0 et 1 et les contrôles réalisés, qui ont permis de choisir la solution mise en oeuvre in fine.

3.5.1 Principe du test et principaux enseignements.

On a comparé deux types de tirage d'UP dans chacune des strates 0 et 1 à savoir :

- un tirage régional systématique selon des probabilités conditionnelles qui assurent in fine des probabilités de tirage des UP EM+EMEX proportionnelles à leur nombre de résidences principales. Dans chaque région, ce tirage a été réalisé selon un tri préalable départemental des UP[12].
- un tirage régional de type Cube, de taille fixe, équilibré au niveau régional sur 1, 2, 3, 4 ou 5 variables selon les régions, en utilisant la méthode de l'équilibrage inverse décrite ci-dessus. Le choix de ces variables se définit dans l'ordre suivant :

  nombre de résidences principales

  revenu net imposable

  nombre de résidences secondaires

  effectif des [20, 59 ans]

[12] Ce tri départemental permet d'assurer la présence d'au moins une UP-EMEX dans chaque département.

effectif des [60,+]

Finalement, l'étude nous a amené à préférer le tirage systématique bien que les simulations montrent que le tirage équilibré donne une meilleure précision en termes d'écarts quadratiques moyens d'estimateurs de totaux relatifs à quelques variables issues du RP99[13] (sauf la variable «revenu net imposable» issue des sources DGI de l'année 1996).

Les raisons de ce choix découlent avant tout de considérations pratiques mais aussi d'un principe d'équité entre les régions sur la précision statistique des extensions régionales (le tirage équilibré conduit, pour certaines régions, à supprimer des variables d'équilibrage, faute de quoi on ne trouverait pas de probabilités d'inclusion admissibles.

Par ailleurs, l'analyse des résultats du tirage systématique exhibe un estimateur du nombre de résidences principales biaisé sur certaines régions, ce qui peut paraître contradictoire compte tenu des probabilités de tirage. La cause du problème provient du tirage initial des UP-EM par la macro Cube qui ne permettait pas de respecter dans certaines régions le nombre théorique d'UP à tirer[14]. Afin de corriger ce défaut, on redresse les poids des UP du tirage systématique (cf. § 3.5.4)

3.5.2 Comparaison du systématique et de CUBE en strate de gestion 0 sur quelques régions.

On a comparé dans le tableau suivant les performances du tirage systématique et du tirage équilibré en strate de gestion 0, sur quelques régions. Les performances sont mesurées par l'EQM et le Biais relatifs qui sont calculés, par simulation, sur 17 variables (dont on connaît les vrais totaux régionaux). Les résultats du tableau sont obtenus par addition des résultats (EQM ou biais relatif) obtenus sur chacune des variables.

Tableau Comparatif.

Région	Nombre de simulations	Variables d'équilibrage	EQM cube	EQM syst.	Biais relatif Cube	Biais relatif Systématique
11	100	1	0.22	0.24	0.06	0.12
21	100	5	0.09	0.56	0.03	0.48
23	100	3	0.09	0.16	0.02	0.11
24	100	5	0.06	0.18	0.02	0.14
25	100	3	0.04	0.22	0.008	0.14
82	100	5	0.08	0.43	0.03	0.27

[13] On se réfère à 17 variables. Ces variables sont les variables d'équilibrage, des effectifs par catégorie de logements (occasionnels et vacants) ainsi que des effectifs par catégorie sociale ou encore des nombres de chômeurs (données RP99).

[14] Pour plus de détails, cf. Vii èmes JMS, 4-5 décembre 2000, « Application de la macro Cube au tirage des UP de l'Echantillon-Maître », L. WILMS, Insee-Méthodes n° 100 (tome 1), pages 154-164.

3.5.3 Analyse des résultats et conclusion.

1) Sur le plan statistique, Cube est incontestablement meilleur : les EQM « sys » et « Cube » obtenues après tirage de 100 échantillons sur quelques régions sont éloquents.

2) Quitte à réduire le nombre de variables d'équilibrage, il est toujours possible de trouver un jeu de probabilités d'inclusion solution du problème de minimisation (ceci a bien été vérifié sur chacune des 22 régions où, dans la majorité des cas, le nombre de variables d'équilibrage oscille entre 4 et 5).

3) Cependant les jeux de probabilités relatifs à l'équilibrage CUBE régional ne permettent pas d'assurer des nombre de FA constant par UP (il suffit, sur une région, qu'une UP ait une « probabilité finale solution » beaucoup plus élevée que la probabilité finale initiale - égale à celle du systématique - pour que cette UP soit tirée et, en conséquence, ait un nombre beaucoup plus faible de FA tirées. Par exemple, sur la région Rhône-Alpes, le nombre de FA par UP oscille de 8 à 40 pour une enquête standard !

On retiendra donc un tirage systématique pour le tirage des UP EMEX de chaque région (on garantit ainsi le même nombre de FA par UP).

3.5.4 Amélioration du tirage systématique (redressement des UP).

On note que le tirage systématique des UP EMEX peut, et de façon paradoxale, engendrer de mauvaises estimations sur le nombre de résidences principales de certaines régions alors que les probabilités de tirage finales (EM+EMEX) sont précisément proportionnelles au nombre de résidences principales.

Ce défaut résulte d'un problème en amont lors du tirage des UP EM et réside dans une faiblesse de la macro de tirage CUBE utilisée à l'époque. Cette faiblesse a été corrigée depuis lors suite au développement d'une nouvelle version de CUBE réalisée par F.TARDIEU ; En l'occurrence, CUBE ne permettait pas de respecter le nombre exact d'UP à tirer par région (variation toutefois le plus souvent nulle, parfois égale à plus ou moins une unité et plus rarement à plus ou moins 2, voire 3 pour PACA). Ainsi, on observe une différence entre le nombre théorique d'UP EM+EMEX à tirer et le nombre d'UP réellement sélectionnées.

Afin de corriger cet effet, on peut redresser les poids de tirage finaux des UP en construisant un estimateur par le ratio calé sur la variable de probabilité d'inclusion finale EM+EMEX (notée p_i). On remarquera que cet estimateur est également calé sur le nombre de résidences PRINcipales de la région (partie rurale).

$$\hat{Y}_R = \frac{\sum_{i \in s_1 \cup s_2} y_i / p_i}{\sum_{i \in s_1 \cup s_2} 1} m_R$$

avec pour l'UP i,

$$p_i = m_R \frac{PRIN_i}{\sum_{i \in Région \cap SGESTION=0} PRIN_i}$$

où m_R désigne le nombre théorique d'UP EM+EMEX attendues.

Pour la mise en oeuvre de ce redressement dans la chaîne de tirage, dans les régions soumises à extension :

1) on calcule les allocations par UP à partir des vraies probabilités d'inclusion finale des UP EM+EMEX (les Π_i). Ceci permet de conserver la propriété d'équiprobabilité du tirage final des logements (« autopondération »).

2) le poids final d'un logement **d'une région à extension** pour une enquête n'est donc plus le poids brut de tirage (i.e., l'inverse de sa probabilité d'inclusion) mais le poids brut de tirage multiplié par le rapport du nombre théorique d'UP EM+EMEX sur le nombre d'UP EM+EMEX réellement tirées.

3.5.5 Qualité[15] du tirage en strate de gestion 0 et 1 avant et après redressement.

A partir de l'échantillon d'UP-EMEX, on mesure dans le tableau suivant la qualité de l'échantillon avant et après redressement. Ces résultats confirment qu'il est impératif d'intégrer dans la chaîne EMEX cette phase de redressement.

Région	**Avant redressement.**		**Après redressement.**	
	Strate 0	**Strate 1**	**Strate 0**	**Strate 1**
11	0.34	0.17	0.34	0.04
21	0.31	0.57	0.01	0.25
22	0.11	0.27	0.07	0.27
23	0.12	0.57	0.12	0.10
24	0.22	0.03	0.04	0.03
25	0.41	0.37	0.21	0.37
26	0.03	0.07	0.03	0.07
31	0.07	0.04	0.07	0.04
41	0.02	0.08	0.02	0.08
42	0.29	0.35	0.04	0.21
43	0.22	0.03	0.11	0.03
52	0.01	0.44	0.01	0.33
53	0.14	0.21	0.07	0.17
54	0.02	1.01	0.03	0.19
72	0.23	0.23	0.19	0.08
73	0.10	0.25	0.11	0.11

[15] La qualité est mesurée par addition des qualités de chacune des 17 variables. La qualité d'une variable est ici définie comme l'écart au carré entre l'estimation du total et le vrai total.

74	0.25	0.12	0.07	0.12
82	**0.63**	0.18	**0.78**	0.10
83	0.09	0.54	0.09	0.54
91	0.22	0.17	0.12	0.12
93	1.84	0.11	0.35	0.01
94	0.12	0.06	0.12	0.06

1. En strate 0, région 82, le redressement est globalement moins bon. Cela est dû à des biais relatifs assez fortement aggravés sur le nombre de résidences occasionnelles et secondaires (une autre façon de dire : mauvaise corrélation sur cette région entre la variable de redressement et le nombre de résidences secondaires ou occasionnelles). Sur les autres variables, l'estimateur redressé est meilleur, voire même nettement meilleur.
2. Si l'on compare les qualités régionales d'un systématique, après redressement, aux EQM régionales CUBE, on constate toujours la supériorité de CUBE.

3.5.6 Considérations pratiques sur le tirage des UP-EMEX.

L'application du tirage CUBE dans le cadre d'hypothèses simplificatrices sur la nature des estimateurs et la forme des probabilités d'inclusion, en utilisant la méthode de l'équilibrage inverse, ne conduit pas nécessairement à des solutions admissibles parce que les probabilités d'inclusion obtenues sont souvent éloignées des probabilités d'inclusion théoriques (proportionnelles à la taille des unités), ce qui risquerait d'altérer la propriété de charge constante par UP.

On a donc finalement recouru à un ***tirage systématique avec probabilités proportionnelles à la taille (en nombre de résidences principales) sur un fichier trié par département***, ce qui réalise une stratification implicite suivant ce dernier critère.

Insistons enfin sur le fait, que *l'EMEX n'est pas construit pour être représentatif au niveau départemental,* même si nous retenons le critère d'une stratification départementale[16] et d'une allocation des échantillons d'Unités primaires proportionnelle à la taille des départements.

Pour les strates de gestion 0 et 1, **l'échantillon total régional d'Unités primaires d'une strate de gestion donnée**, assure une « représentativité » régionale dans les conditions suivantes : :

- au moins une Unité primaire de cet échantillon tombe dans chaque département[17]

[16] Stratification départementale ne veut pas dire représentativité départementale mais seulement assurance que l'extension régionale couvrira correctement l'ensemble des départements.

[17] On pourrait aussi, dans le souci d'assurer une meilleure dispersion géographique et de satisfaire des demandes potentielles, imposer *que le chef-lieu de chaque département soit retenu d'office dans l'EMEX.*

- il y a une stratification départementale implicite : l'échantillon d'Unités primaires se répartit par département au prorata du nombre total d'Unités primaires constituées dans chaque département.

4 Tirage des logements à l'intérieur des Unités primaires : groupes de communes et districts.

- La constitution des groupes de communes (quelle que soit la strate de gestion) s'effectuera sur les mêmes critères que ceux ayant prévalu pour l'EM national.
- En strate 2, comme pour l'EM, un tirage de groupes de districts équilibré au niveau de chaque UP-EMEX est réalisé (mais il n'y a pas de tirage de nouveau district dans les UP-EM).
- En strate 3, on réalise un tirage de districts EMEX, de telle sorte que l'ensemble des districts EM et EMEX soit équilibré sur chaque Unité urbaine.
- En strate 4, la procédure est similaire à celle de la strate 3, mais avec équilibrage départemental.

4.1 Constitution des groupes de communes (GRCOM).

On partitionne chaque UP rurale et chaque UU de l'EMEX en GRCOM. Les GRCOM-EMEX jouent le même rôle que ceux de l'EM et obéissent aux mêmes règles de constitution :

1) Pour le rural, les UP étant relativement étendues géographiquement, dans le cas d'une enquête avec extension, les GRCOM des UP-EMEX rurales seront échantillonnées de façon à limiter les déplacements de l'enquêteur au sein de l'UP.
2) Pour les autres strates de gestion, comme les UU ne posent pas de réels problèmes de déplacement, le rôle des GRCOM est à l'opposé : tous les GRCOM sont impactés lors d'une enquête si bien que l'on évite tout phénomène de concentration géographique des FA à enquêter dans l'UU.

La constitution des GRCOM a été réalisée de façon automatisée et en ré-utilisant les mêmes algorithmes que ceux employés pour l'EM, sauf sur les UP-EMEX rurales.

En effet, pour ces dernières, on a utilisé un algorithme amélioré, développé par M. CHRISTINE et M. ISNARD[18], qui a permis d'intégrer et de respecter l'ensemble des contraintes suivantes :

- au sein d'un GRCOM, les communes sont contiguës
- la taille minimum d'un GRCOM est contrôlée (100 logements)

[18] VII èmes JMS, 4-5 décembre 2000, « Un algorithme de regroupement d'unités statistiques selon certains critères de similitude », Insee-Méthodes n°101 (tome 2), pages 71-120.

- on privilégie, au sein d'un GRCOM, des associations de petites et grandes communes (critère d'hétérogénéité des tailles des communes)
- on évite la formation de GRCOM «trop gros » en contrôlant le nombre de GRCOM à créer dans chaque UP.

Exemples cartographiques[19] de GRCOM créés dans des UP rurales : ces exemples sont illustrés dans un document séparé.

On notera enfin qu'il est inutile de créer des GRCOM-EMEX spécifiques sur les UU des strates 3 et 4 : ils ont déjà été créés lors de la constitution de l'EM.

4.2 Tirage des districts en strates de gestion 2, 3 et 4.

Comme pour la constitution de l'EM, le tirage des districts-EMEX vise à réduire le volume de logements à stocker tout en garantissant une réserve suffisante pour les enquêtes à extensions prévues (cf. § 4.4 pour plus de détails sur l'aspect volumétrie).

De surcroît, le tirage des districts permet de résoudre la question de l'articulation avec les enquêtes purement locales : pour assurer une disjonction entre échantillons EMEX et échantillons d'initiative locale, on peut retenir comme principe que ces derniers seront puisés hors districts EMEX (et hors districts EM).

Les districts sont sélectionnés selon un tirage équilibré au moyen de la macro Cube. Les variables d'équilibrage sont :

- nombre de résidences principales
- revenu net imposable
- effectif des [20, 59 ans]
- effectif des [60,+]
- nombre de résidences secondaires.

4.2.1 Tirage des groupes de districts en strate 2.

Le tirage des groupes de districts-EMEX n'est effectué qu'au sein des UP-EMEX. Il se définit de façon identique à celui des districts-EM dans les UP-EM.

Dans chaque GRCOM, les districts des UP-EMEX sont éventuellement associés en groupes de districts de 100 logements au minimum. Puis, 30 groupes de districts sont alors sélectionnés à probabilités égales selon un tirage équilibré au niveau de l'UP et stratifié par GRCOM avec allocation proportionnelle.

On notera que ce type de tirage équilibré n'entre pas dans la problématique du tirage équilibré conditionnel soulevé au § 3. En effet, dans ces UP-EMEX, il n'y a pas eu de tirages préalables

[19] La cartographie a été réalisée par G. BOURDALLE.

de groupes de districts-EM. Il n'y a donc pas de difficulté pour atteindre la propriété d'équilibrage.

4.2.2 Tirage des districts en strate 3 et 4.

Contrairement à la strate de gestion 2, le tirage des districts représente un véritable premier degré de tirage.

Bien que les tirages se définissent, dans chaque UU, conditionnellement à celui des districts-EM, il est toutefois possible d'obtenir la propriété d'équilibrage sur un ensemble de variables au niveau de chaque UU à partir de l'échantillon {districts-EM, districts-EMEX}. Précisons que, pour l'UU de Paris, la propriété d'équilibrage est assurée au niveau de chacun des départements composant cette UU.

Cette propriété favorable est en effet assurée grâce aux tirages *à probabilités égales* des districts EM et des districts-EMEX (cf. annexe 4 pour la démonstration de ce résultat).

Le tirage assure un nombre de districts-EMEX quasi-égal au nombre de districts-EM dans chacun des GRCOM, soit 5% des districts, de façon que, pour une extension régionale standard, 50% des logements proviennent de l'EM et 50 % de l'EMEX.

On remarquera qu'il y a, au sein de chacun des GRCOM, coexistence entre les logements EM et EMEX. Ce n'est pas le cas pour les GRCOM des autres strates de gestion.

4.3 Constitution définitive de la base EMEX et bilan.

On charge tous les logements RP99 appartenant aux UP-EMEX des strates de gestion 0 et 1 et l'on charge tous les logements RP99 appartenant aux districts-EMEX des strates de gestion 2, 3 et 4.

On achève la constitution de la base EMEX, partie logements recensés, après suppression des logements de l'Enquête Emploi en Continu présents dans l'EMEX (cette disjonction est nécessaire afin d'éviter de solliciter un même ménage pour ces deux enquêtes).

Au final :

Les UP-EMEX sont issues de la même base de sondage que celle des UP-EM. L'effort de constitution de la base de sondage des UP-EMEX était donc faible et s'est principalement limité à l'exclusion des UP-EM de la base initiale et à l'adjonction des probabilités de tirage des UP-EMEX.

En strates de gestion 0, 1 et 2, le tirage des UP-EMEX, qui s'effectue dans la base de sondage des UP privée des UP-EM, conduit à définir de nouvelles zones géographiques d'enquêtes a priori éloignées des zones EM et exigeront donc, le plus souvent, à recruter de nouveaux enquêteurs en cas d'enquêtes à extensions (cf. cartographie).

En strates 3 et 4, toutes les UU sont déjà dans l'EM et demeurent dans l'EMEX (UU mixte). Le tirage des districts-EMEX se déroule donc dans chacune des UU de plus de 100 000 habitants et conduit à définir les zones d'enquêtes complémentaires en cas d'extension (cf. cartographie).

La constitution et la gestion de la base des logements neufs dans les UP-EMEX sont traitées au paragraphe 7.

4.4 Eléments de volumétrie des bases de logements EM et EMEX.

Toutes strates de gestion confondues, le nombre de logements dans la base EMEX (après disjonction de l'échantillon Emploi) est de : contre 2 000 000 de logements dans la base EM.

On peut également retenir que la taille de la base EM+EMEX est, par rapport à celle de la base EM :

b) le double en strate de gestion 3 et 4

c) un peu moins du double en strate de gestion 2, car le taux de sondage apparent des UP-EM est déjà de 93/187

d) le triple en strate de gestion 0 et 1, car le nombre d'UP EMEX est le double du nombre d'UP-EM.

L'annexe 5 donne plus de détails sur la volumétrie des deux bases.

5 La question des logements neufs.

Tout ce qui a été développé dans les paragraphes qui précèdent est relatif au tirage des logements RP. Or il est évidemment essentiel de pouvoir introduire des logements « neufs » (i.e., construits après le RP) pour les extensions régionales d'enquête.

5.1 Les principes généraux.

Motivations :

- Il faut une base de sondage de logements neufs.
- On ne peut pas se permettre de « décalquer » le système BSLN[20] pour la partie EMEX, ce serait trop coûteux d'effectuer un suivi complémentaire de la construction dans chaque région sans une garantie suffisante d'utilisation effective de la base.
- On dispose déjà du fichier SITADEL, contenant l'information souhaitée, France entière, chaque trimestre.

[20] Base de sondage des logements neufs, fondée sur un tirage de logements dans les fichiers de permis de construire et un suivi annuel par enquêteur pour valider l'état d'achèvement des logements.

Le suivi sur le terrain d'une base de sondage de logements neufs hors échantillon-maître est peu envisageable (moyens considérables à dégager et planifier). En outre, toute région devrait suivre une telle base en continu dans l'éventualité d'une extension qui pourrait très bien ne jamais avoir lieu.

Une alternative serait de tirer directement un échantillon de logements neufs dans les fichiers de permis de construire, mais avec un taux de chute inconnu, puisqu*'on ne peut tirer que dans les permis mis en chantier*, donc non nécessairement achevés à la date de l'enquête. Pour les enquêtes d'un futur proche, on a cependant l'avantage de ne pas être trop loin du RP et de connaître les DR sujettes à extension : il serait envisageable, pour celles-ci, de dégager des moyens spécifiques pour constituer une BSLN « EX ». Mais des inégalités de traitement apparaîtraient vis-à-vis des autres régions.

On peut penser aussi à d'autres fichiers : fichiers fiscaux (on pense en premier lieu à la taxe foncière), EDF, compteurs d'eau, etc… mais on ne dispose pas de ces fichiers actuellement. Il faudrait de surcroît s'assurer que ces fichiers contiennent bien l'information qui nous est nécessaire pour vérifier la «qualité » de telles sources et pour gérer toutes les questions informatiques (en particulier adaptation des formats de fichiers et normalisation des adresses), les aspects financiers éventuels et les conventions. C'est une piste intéressante, mais à plus long terme.

Il a donc été décidé, ***au moins dans un premier temps, d'utiliser comme base de sondage la liste des logements des permis déclarés <u>achevés</u> dans SITADEL. On tirerait alors un supplément de logements neufs dans les UP « EMEX ».***

<u>Risques</u> :

- Celui du défaut d'exhaustivité, lié à 2 phénomènes :
  - Logements achevés non repérés comme tels par Sitadel.
  - Logements achevés finissant par être pris en compte, mais tardivement (délai entre date réelle d'achèvement et date de prise en compte par le système informatique).

  Une réunion avec le Ministère de l'Equipement a permis d'éclaircir la situation : on peut estimer que, pour 100 permis achevés l'année n, il y en a un peu moins de 20% qui, pour toutes ces raisons, ne sont pas dans le fichier de l'année n+1 (parmi lesquels 75% ne le seront jamais, pour des raisons que nous ignorons).

- Difficultés éventuelles pour retrouver l'adresse. On peut craindre que certaines adresses attribuées au moment de l'avis favorable soient en fait provisoires (type adresses de chantier, ou du maître d'ouvrage), et permettent d'identifier un immeuble durant sa construction, mais pas nécessairement plusieurs années après. Il faudrait effectuer un test de pertinence des adresses.

  Un îlotage automatique est cependant envisagé : la connaissance du taux d'échec de l'appariement sur les adresses donnera une première approche de la qualité des adresses dans ce fichier, au moins dans le gros urbain.

<u>Conclusion</u> : La base est incomplète certes, mais il est préférable d'avoir une base incomplète que pas de base du tout.

<u>Un problème général</u> : se pose, dans tous les cas, l'épineux problème de l'évaluation du volume de construction neuve au niveau régional - sachant qu'au niveau national, les estimations que l'on pratique actuellement sont entachées d'imprécision.

Concrètement, pour les régions à extension, un premier tirage « à blanc » de la chaîne EM permet de déterminer les allocations en logements neufs et en logements RP, résultant de l'évaluation nationale de la construction neuve et de la taille de la BSLN régionale. Sur la base d'une extension dont le volume est en général fixé par les commanditaires en supplément d'échantillon, le partage entre logements neufs et logements RP de ce supplément est calculé au prorata des allocations calculées lors du tirage à blanc.

⇨ Au final :

La proposition d'utilisation de la partie « logements achevés » de SITADEL a été adoptée. Le défaut d'exhaustivité constitue un risque réel mais il ressort que si celui-ci touche de manière indifférenciée l'ensemble des types de logements neufs, alors le biais lié à ce défaut de couverture sera négligeable (une étude sera menée ultérieurement à partir d'un rapprochement entre la BSLN et ce fichier des logements déclarés achevés).

Si des problème sérieux de qualité étaient à l'avenir mis en évidence - ce qu'on ne peut exclure - l'alternative pourrait être d'engager en région une opération spécifique de suivi de construction sur le terrain pour compléter localement la base BSLN dans les UP de l'EMEX. Les partenaires locaux seraient alors confrontés à la nécessité d'un financement spécifique motivé par un argument de qualité. Ce suivi serait confié aux Directions régionales.

5.2 Mise en oeuvre.

La réserve de logements neufs dans laquelle on va puiser au moment du tirage des logements neufs pour une enquête donnée sera donc composée :

- pour les régions non soumises à extension : de la BSLN ordinaire
- pour les régions soumises à extension, il faut distinguer suivant les strates de gestion.
    - En strates de gestion 0, 1 et 2, on retient d'une part la BSLN ordinaire dans les UP-EM, d'autre part, la totalité des logements issus des fichiers SITADEL décrits ci-dessus, pour les UP-EMEX.
    - En strates de gestion 3 et 4, la BSLN ordinaire est complétée par un tirage au taux de 1/20 dans les fichiers SITADEL, au sein de chaque Unité urbaine (ce qui assure une réserve de taille équivalente à celle disponible pour le seul EM).

La chaîne de tirage intégré puise indifféremment dans ces deux réserves pour nourrir l'échantillon de logements neufs. On peut donc ré-utiliser la chaîne de tirage existante avec un minimum d'adaptations de programmes.

6 Chaîne de tirage informatique des enquêtes à extension.

Le temps effectif de constitution de la base de logements EMEX, opération assez complexe, a été pourtant relativement court dès lors que les principes méthodologiques de son tirage ont été clairement établis. En effet, cette opération bénéficiait d'une capitalisation d'expérience importante au sein de l'équipe de projet qui venait de constituer l'EM99. Ainsi, on peut estimer la durée de développement du projet EMEX à 8 mois (constitution de la base EMEX et adaptation de la chaîne EM au tirage d'enquêtes avec extension) alors que le développement de l'EM a duré un an et demi avec les mêmes moyens humains.

6.1 Fonctionnalités.

La chaîne EMEX de tirage d'enquêtes est conçue comme une généralisation de la chaîne EM[21]. Ainsi l'utilisateur ne dispose que d'une unique application de tirage d'enquêtes et retrouve tous les paramètres et options de tirages classiques de la chaîne EM, que l'enquête soit avec ou sans extensions, à savoir essentiellement :

- nombre de logements RP et neufs à tirer
- possibilité de sous-représentation des logements par catégorie (principale, secondaire, occasionnelle, vacante, neuve)
- possibilité de sous représentation des logements par strate de gestion
- tirage en deux phases permettant de sur-représenter tel ou tel type de ménage
- paramétrage de la charge minimale de collecte par UP (strates 0 et 1)
- tirage d'enquêtes à vague (découpage d'un échantillon en vagues de collecte de même taille)
- tirage d'individus Kish

Dans le cas d'une région bénéficiant d'une extension, la chaîne EMEX tire des logements RP et des logements neufs dans les UP-EM et dans les UP-EMEX des strates de gestion 0,1 et 2 (cf. en infra, la constitution de la base de sondage des logements neufs sur les UP-EMEX). Pour les unités urbaines des strates de gestion 3 et 4, des districts EMEX complètent les districts EM et ne sont sollicités qu'en cas d'extension sur la région. De même, la BSLN-EM des UU des strates de gestions 3 et 4 est complétée par une base de logements neufs. Cette base complémentaire de logements neufs est, comme pour la BSLN, enrichie annuellement par les logements construits l'année précédente.

Pour le responsable de tirage d'enquête, les seuls paramètres nouveaux à entrer dans la chaîne dans le cas d'une enquête avec extensions sont :

- les régions concernées par l'extension
- les nombres de logements à tirer dans les régions à extension.

En sortie de chaîne, le responsable récupère un fichier classique de tirage (le « FICFLCLE ») contenant notamment les identifiants des logements tirés ainsi que leur poids de tirage. Par ailleurs, l'impression automatisée des fiches-adresse se déroule de la même façon que pour une enquête sans extension.

6.2 Principe de fonctionnement.

[21] La chaîne EM est destinée à tirer les enquêtes dans les UP-EM seulement et n'a donc vocation qu'à tirer les enquêtes nationales ne bénéficiant d'aucune extension.

Très schématiquement, la chaîne de tirage EMEX peut se décrire comme étant composée d'un moteur principal remplissant les deux fonctions suivantes :

- détermination des nombres (allocations) de logements à tirer dans chaque GRCOM
- tirage effectif des logements respectant les allocations dans chacun des GRCOM.

Ce moteur peut fonctionner à partir de n'importe quelle liste de GRCOM d'enquêtes. La chaîne EMEX consiste ainsi, pour l'essentiel, à utiliser le moteur de la chaîne EM et à ajouter un module amont de préparation de la liste des GRCOM à impacter selon la région. Ainsi, pour un tirage donné, dans une région sans extension, la liste en entrée correspond aux GRCOM-EM seuls avec leur probabilité de tirage. Dans une région avec extension, la liste correspond aux GRCOM-EM et aux GRCOM-EMEX avec leur probabilité de tirage finale tenant compte du tirage de l'EMEX.

En fonction des paramètres de tirage et des probabilités finales de sélection des GRCOM, la chaîne calcule les allocations de logements dans chacun des GRCOM.

Elle effectue ensuite la sélection des logements dans les GRCOM selon un tirage systématique (même mode de tirage que celui de la chaîne EM) mais en prenant soin de faire le tirage soit dans la base des logements EM seule, soit dans la base des logements EM+EMEX selon que le GRCOM appartient ou non à une région bénéficiant d'une extension.

Après le tirage, la chaîne marque les logements sélectionnées afin d'éviter de les retirer lors d'une enquête ultérieure (« dogme » de la non-réinterrogation d'un même ménage). Ce marquage des logements tirés permet également de gérer la réserve de logements EM en cas d'épuisement de GRCOM : la chaîne répartit l'allocation en logements à tirer d'un GRCOM épuisé entre les autres GRCOM de l'UP.

7 Exploitation conjointe des données issues de l'EM et de l'EMEX.

La phase d'échantillonnage a été décrite de manière très minutieuse car c'est elle qui fonde la spécificité de l'EMEX. On sera beaucoup plus allusif sur les autres aspects du traitement aval.

L'examen de l'échantillonnage a permis aussi de mettre en lumière les calculs d'estimateurs utilisant toute l'information, nationale et provenant de l'extension. Il s'agit ici d'estimateurs fondés sur les pondérations initiales, calculées au niveau des unités primaires de l'EM et des UP de l'EMEX dans le cadre du tirage conditionnel.

<u>Les questions qui se posent en aval.</u>

a) la correction de la non-réponse totale.

Une correction par estimation de la probabilité de non-réponse nécessitera sans doute de traiter séparément les régions à extension des autres. En particulier, si l'on admet que les extensions sont souvent légitimées par le désir de mettre en évidence des particularités régionales ou des effets propres à la région, il paraît souhaitable que les facteurs explicatifs de la non-réponse, d'une part soient analysés séparément dans les régions à extension et les autres, d'autre part soient même si nécessaire spécifiques au sein des régions à extension.

b) calage, redressement.

Là aussi, les critères mis en oeuvre peuvent être différents dans les régions à extensions et, au minimum, doivent être écrites séparément pour ces régions et les autres.

c) correction de la non-réponse partielle.

Des techniques par imputation, par exemple avec recherche de «donneurs », doivent aussi distinguer les régions à extension des autres : on peut en effet admettre qu'un non-répondant d'une région standard soit remplacé par les réponse d'un répondant d'une autre région standard, mais, pour les régions à extension, il peut sembler préférable que l'imputation se fasse au sein de la même région.

d) calcul de précision.

Il pourra être mené de manière globale dans la mesure où l'on dispose d'estimateurs utilisant l'ensemble de l'information et servant aussi bien pour les estimations nationales que régionales.

L'ensemble de ces procédures devront être testées, mises au point et appliquées au fur et à mesure que l'exploitation de l'enquête Santé avancera.

8 Conclusion.

L'outil EMEX a été conçu principalement pour répondre à des demandes régulières des régions, même si celles-ci sont en faible nombre chaque année.

Mis en place pour la 1ère fois en 2002, l'EMEX est encore un outil trop « jeune » pour qu'on puisse déjà en tirer un bilan. Néanmoins, on peut juger en particulier des qualités de cet outil en le comparant aux modes de gestion antérieurs des extensions régionales et, notamment, de celle de l'enquête Logement : celle-ci a eu lieu en 2001 mais n'avait pas pu bénéficier des fonctionnalités de l'EMEX, dont les développements n'étaient pas achevés à cette date.

Or, il apparaît clairement, lorsque l'on interroge les responsables, que des difficultés nombreuses sont apparues pour cette extension non bénéficiaire de l'EMEX, notamment pour le calcul des poids à intégrer dans la perspective d'une exploitation conjointe de l'enquête nationale et de l'extension.

A l'inverse, si l'on interroge les responsables de l'enquête Santé, à laquelle s'est appliqué l'EMEX pour la première fois, des demandes fortes émergent :

- souhait que les utilisateurs puissent utiliser l'ensemble du fichier pondéré (national + extension, soit 25000 FA) de façon qu'il soit «représentatif » de l'ensemble de la population française, et aussi que l'échantillon de chaque DR à extension puisse être représentatif de la région.
- souhait de pouvoir comparer ses résultats, soit à l'ensemble de la France, soit aux autres régions à extension.

- souhait d'avoir une seule pondération finale.
- souhait d'avoir un traitement coordonné de la non-réponse et surtout des réponses partielles (individus n'ayant pas répondu aux trois visites et abandonné le questionnement en cours d'enquête).

En termes d'organisation de la collecte, le système EMEX présent l'avantage de procéder comme l'échantillon-maître : le tirage des UP-EMEX est fait une fois pour toutes pour toute la durée de vie de l'EMEX (et de l'EM) : les directions régionales savent donc quels sont les lieux où tomberont d'éventuelles extensions régionales et cela peut leur faciliter le travail de recrutement des enquêteurs.

En termes d'organisation informatique, ces similitudes ont permis de réutiliser la chaîne de tirage mise au point pour l'EM, même s'il est vrai que l'intégration dans un module commun du tirage de l'EM et de l'EMEX pour les régions concernées a nécessité des développements conséquents. Mais, pour l'utilisateur final de la chaîne, le système présente une automaticité et une transparence bénéfiques.

L'ensemble des souhaits semblent, à l'heure actuelle, largement couverts par la démarche suivie, même s'ils n'ont pas tous encore été mis en oeuvre (notamment, tous les aspects aval du traitement).

En particulier, il n'y aura bien qu'une seule pondération finale, calculée en sortie de la chaîne de tirage, pour fournir des estimations sans biais à tous niveaux. Ainsi, quelle que soit la zone géographique, l'utilisation du poids donne un estimateur sans biais - c'est la théorie des estimations dite "par domaine". Par exemple, pour estimer sur la ville de Paris, on retient seulement les logements de Paris et on utilise "le" poids du fichier national.

Naturellement, se pose après un problème de précision, donc de pertinence du niveau de détail possible pour la ventilation des résultats.

Il conviendra de juger, à l'aune de cette première application et une fois les traitements aval mis en œuvre, si la satisfaction des « clients » est suffisante : ainsi, à ce prix, pourra-t-on juger de la pleine capacité de l'outil à répondre à la demande et il ne fait nul doute que cela sera certainement un critère déterminant pour susciter de nouvelles demandes.

ANNEXE 1 :
Liste des UU de la strate de gestion 2
en dehors de l'EM.

LIBELLE	D	UU99	RGFORCE
BEAUMONT-SUR-OISE	60	00456	11
CHAMPAGNE-SUR-SEINE	77	77404	11
COULOMMIERS	77	77403	11
MONTEREAU-FAULT-YONNE	77	77405	11
ESBLY	77	77402	11
OZOIR-LA-FERRIERE	77	77401	11
ETAMPES	91	91401	11
SEDAN	08	08401	21
EPERNAY	51	51401	21
CHATEAU-THIERRY	02	02401	22
LAON	02	02403	22
BEAUVAIS	60	60501	22
CHANTILLY	60	60401	22
COMPIEGNE	60	60502	22
LOUVIERS	27	27402	23
EVREUX	27	27501	23
BARENTIN	76	76401	23
DIEPPE	76	76403	23
FECAMP	76	76402	23
EU	76	00451	23
VIERZON	18	18401	24
VENDOME	41	41401	24
TROUVILLE-SUR-MER	14	14401	25
SAINT-LO	50	50401	25
ALENCON	61	00461	25
MACON	01	00463	26
CREUSOT	71	71401	26
AUXERRE	89	89402	26

CAMBRAI	59	59402	31
ARMENTIERES (*)	59	59501	31
HAZEBROUCK	59	59401	31
BERCK	62	62401	31
SAINT-OMER	62	62501	31
LUNEVILLE	54	54402	41
PONT-A-MOUSSON	54	54403	41
BAR-LE-DUC	55	55401	41
SAINT-AVOLD (*)	57	57402	41
SARREGUEMINES (*)	57	57401	41
SAINT-DIE	88	88402	41
EPINAL	88	88501	41
HAGUENAU	67	67501	42
GUEBWILLER	68	68401	42
PONTARLIER	25	25401	43
LONS-LE-SAUNIER	39	39401	43
BELFORT	70	00556	43
VESOUL	70	70401	43
ROCHE-SUR-YON	85	85402	52
DINAN	22	22401	53
DINARD	22	00452	53
CONCARNEAU	29	29402	53
MORLAIX	29	29403	53
PENMARCH	29	29401	53
FOUGERES	35	35401	53
COGNAC	16	16401	54
ROYAN	17	17402	54
SAINTES	17	17401	54
CHATELLERAULT	86	86401	54
LIBOURNE	33	33401	72
DAX	40	40402	72
MONT-DE-MARSAN	40	40401	72
MARMANDE	47	47401	72
MILLAU	12	12401	73
RODEZ	12	12402	73

MAZAMET	81	81401	73
CHARVIEU-CHAVAGNEUX	01	00453	82
MONTELIMAR	07	00460	82
ANNONAY	07	07402	82
AUBENAS	07	07401	82
VIENNE	38	00462	82
VILLEFONTAINE	38	38401	82
VOIRON	38	38403	82
SAINT-JUST-SAINT-RAMBERT	42	42401	82
AIX-LES-BAINS	73	73402	82
ALBERTVILLE	73	73401	82
THONON-LES-BAINS	74	74502	82
RIOM	63	63401	83
CARCASSONNE	11	11401	91
NARBONNE	11	11402	91
ALES	30	30501	91
LATTES	34	34401	91
LUNEL	34	34402	91
MANOSQUE	04	04401	93
GAP	05	05401	93
MIRAMAS	13	13401	93
ISTRES	13	13402	93
DRAGUIGNAN	83	83401	93
ISLE-SUR-LA-SORGUE	84	84401	93

ANNEXE 2 : équilibrage ex-ante

de l'ensemble EM + EMEX.

Si la perspective EMEX avait été prise en compte au moment du tirage de l'EM, il y aurait alors eu un moyen très simple d'équilibrer les 2 échantillons :

→ Tirer un échantillon S de k unités dans ***P*** avec des probabilités Π_i en équilibrant sur X :

$$\sum_{i \in S} \frac{X_i}{\Pi_i} = \sum_{i \in P} X_i \ .$$

Cela donnait EM + EMEX

→ Enlever de S, par sondage aléatoire simple au taux f équilibré sur la variable

$Z_i = \frac{X_i}{\Pi_i}$ un échantillon S' de taille fixe (k - a) :

$$\sum_{i \in S'} \frac{X_i}{f\,\Pi_i} = \sum_{i \in S} \frac{X_i}{\Pi_i}.$$

L'échantillon S - S' serait alors tiré de manière à assurer à chaque unité i la probabilité de sélection Π_i^1 = (1 - f) Π_i. Si Π_i est proportionnelle à la taille de l'unité i, il en va de même de Π_i^1.

On vérifie enfin facilement que :

$$\sum_{i \in S-S'} \frac{X_i}{(1-f)\Pi_i} = \sum_{i \in P} X_i \ ,$$

c'est-à-dire que S - S' est également équilibré.

ANNEXE 3 : une approche des propriétés des « échantillonnages successifs ».

Considérons une population ***P*** de taille N, dont les éléments sont dénommés "unités". A chaque unité i de la population est associée une variable d'intérêt Y_i .

On tire dans cette population un premier échantillon sans remise, s_1. Puis, conditionnellement à la réalisation de ce premier tirage, on tire un second échantillon sans remise, s_2.

On cherche alors à construire des estimateurs sans biais du total sur la population $T = \sum_{i \in P} Y_i$ à partir de ces deux échantillons.

1. Rappel des notations.

Un échantillon s sans remise peut être assimilé à un vecteur de $\boldsymbol{R}^N$ dont la i^e composante $\underline{S}_i$ vaut 1 si l'unité i appartient à s, 0 sinon. Tirer aléatoirement un échantillon revient à définir une loi de probabilité sur $\{0, 1\}^N$.

Un échantillon s est alors la réalisation d'une variable aléatoire S à valeurs dans $\{0, 1\}^N$. Sa loi est définie par les valeurs P {S = s} pour tout $s \in \{0, 1\}^N$.

La probabilité d'inclusion d'ordre 1 d'une unité i, notée Π_i, est alors définie par :

Π_i = P {i ∈ S}

$$= \sum_s P(S = s \text{ et } i \in s)$$

$$= \sum_{s / i \in s} P(S = s) .$$

De même, on peut écrire :

$$1 - \Pi_i = \sum_{s / i \notin s} P(S = s) .$$

Dans la suite, on désignera de la même façon l'échantillon S en tant que partie aléatoire de ***P*** ou en tant que vecteur aléatoire dans $\{0, 1\}^N$ et on notera i ∈ S pour exprimer que la i^e composante de S, soit $\underline{S}_i$, vaut 1, c'est-à-dire que l'unité i est dans l'échantillon S (*on rappelle qu'en théorie classique des sondages, c'est S qui est une variable aléatoire et non i*).

On aura alors : $\Pi_i = \sum_s P(S = s \text{ et } \underline{S}_i = 1)$

$$= \sum_s P(S = s \text{ et } \underline{s}_i = 1)$$

$= \sum_s \underline{s}_i \; P(S = s)$

$= E\, \underline{S}_i$

2. Mode de constitution des deux échantillons et lois.

Le mode de constitution est défini par les lois de tirage des deux échantillons :

- Pour le 1er échantillon, une loi définie par $\boldsymbol{P}\{S_1 = s_1\}$ pour $s_1 \in \{0, 1\}^N$ et qui est résumée par la probabilité d'inclusion d'ordre 1 de chaque unité i appartenant à P :

$$\Pi_i^1 = P\{i \in s_1\}$$

Exemple : le 1er échantillon, de taille fixe n_1 est tiré de manière équiprobable sans remise. On a alors : $P\{S_1 = s_1\} = \frac{1}{C_N^{n_1}}$ et $\Pi_i^1 = \frac{n_1}{N}$.

- Pour le 2ème échantillon, tiré postérieurement au 1er, sa constitution est définie par une loi conditionnelle aux valeurs du 1er échantillon, notée mathématiquement $P_{(s_2)}^{S_2/S_1=s_1}$ ou $P(S_2 = s_2 / S_1 = s_1)$.

  Cette loi est elle-même résumée par les probabilités d'inclusion conditionnelles d'ordre 1 des unités i, notées :

  $\Pi_i^{2/S_1=s_1} = P(i \in S_2 / S_1 = s_1)$, ou, plus simplement, Π_i^{2/s_1}.

  Il s'agit d'une fonction de s_1, soit h (s_1). La variable aléatoire correspondante h (S_1) sera notée plus simplement Π_i^{2/S_1}.

Ces deux lois de tirage sont les seules données de base.

Remarquons à ce stade qu'il n'est pas utile de faire d'hypothèse particulière sur les relations entre les deux échantillons : œux-ci peuvent être totalement indépendants [ce qui implique alors $? \; {}_i^{2/s_1} = ? \; {}_i^{2} \; \forall s_1$], ou bien au contraire, le 2ème échantillonnage dépendra du premier.

Les cas de non indépendance se rencontrent en particulier lorsque le champ du tirage du 2ème échantillonnage est défini à partir du résultat du 1er. Deux cas usuels illustrent cette éventualité :

a) celui où le 2ème échantillon est tiré dans le complémentaire du 1er. Ceci correspond à la pratique courante en matière d'échantillonnage des enquêtes « ménages » : le principe de non-réinterrogation entraîne que tout échantillon d'enquête est tiré dans le complémentaire des échantillons tirés antérieurement au cours de la même période inter-censitaire. De même, dans le cas de l'EMEX (pour les strates de gestion 0 et 1), on tirera les unités primaires de celui-ci dans le complément de celles déjà tirées pour l'échantillon-maître national.

b) à l'opposé, celui où le 2ème est inclus dans le 1er : c'est le cas de *l'échantillonnage en deux phases.*

Exemple : le 2ème échantillon, de taille fixe n_2, est tiré de manière équiprobable sans remise dans le complémentaire de S_1. On a alors :

$P\{S_2 = s_2 / S_1 = s_1\} = 0$ si $s_1 \cap s_2 \neq \emptyset$

$$= \frac{1}{C_{N-n_1}^{n_2}} \text{ si } s_1 \cap s_2 = \emptyset.$$

Et :

$?_i^{2/S_1=s_1} = 0$ si $i \in s_1$

$$= \frac{n_2}{N-n_1} \text{ sinon.}$$

Loi conjointe des deux échantillons.

$P\{S_1=s_1 \text{ et } S_2=s_2\} = P\{S_2=s_2 / S_1=s_1\} P\{S_1=s_1\}$.

Réunion des deux échantillons.

On peut calculer la probabilité d'inclusion finale d'une unité i dans la réunion des deux échantillons :

$$\begin{aligned}\Pi_i &= P(i \in S_1 \cup S_2) \\ &= P(i \in S_1) + P(i \in S_2) - P(i \in S_1 \cap S_2) \\ &= P(i \in S_1) + P(i \in S_2 \cap \overline{S}_i) \\ &= P(i \in S_1) + P(i \in S_2 / i \notin S_1) P(i \notin S_1)\end{aligned}$$

soit :

$$\Pi_i = \Pi_i^1 + P(i \in S_2 / i \notin S_1)(1 - \Pi_i^1)$$

Calcul de la probabilité conditionnelle $P(i \in S_2 / i \in S_1)$ ou $P(i \in S_2 / i \notin S_1)$

Il importe de bien distinguer cette probabilité conditionnelle des probabilités Π_i^{2/s_1}.

On a :

$$P(i \in S_2 / i \notin S_1) = \frac{P(i \in S_2 \text{ et } i \notin S_1)}{P(i \notin S_1)}$$

$$= \frac{\sum_{s_2} (i \in S_2 \text{ et } i \notin S_1 \text{ et } S_2 = s_2)}{P(i \notin S_1)}$$

$$= \frac{\sum_{s_1}\sum_{s_2} \left(i \in s_2 \text{ et } i \notin s_1 \text{ et } S_2 = s_2 \text{ et } S_1 = s_1\right)}{P\left(i \notin S_1\right)}$$

$$= \sum_{s_1 / i \notin s_1} \frac{\sum_{s_2} P\left(i \in s_2 \text{ et } S_2 = s_2 \text{ et } S_1 = s_1\right)}{P\left(i \notin S_1\right)}$$

$$= \sum_{s_1 / i \notin s_1} \frac{P(i \in S_2 \, et \, S_1 = s_1)}{P(i \notin S_1)}$$

$$= \sum_{s_1 / i \notin s_1} \frac{P\left(S_1 = s_1\right) P\left(i \in S_2 / S_1 = s_1\right)}{P\left(i \notin S_1\right)}$$

$$= \frac{\sum_{s_1 / i \notin s_1} \Pi_i^{2/s_1} P\left(S_1 = s_1\right)}{P\left(i \notin S_1\right)}$$

D'où :

$$\boxed{P\left[i \in S_2 / i \notin S_1\right] = \frac{\sum_{s_1 / i \notin s_1} \Pi_i^{2/s_1} P\left(S_1 = s_1\right)}{1 - \Pi_i^1}}$$

En utilisant la relation démontrée ci-dessus, on obtient donc :

$$\boxed{? _i = \Pi_i^1 + \sum_{s_1 / i \notin s_1} \Pi_i^{2/s_1} P\left(S_1 = s_1\right)}$$

Ou encore :

$$\Pi_i = \Pi_i^1 + \sum_{s_1} \Pi_i^{2/s_1} 1_{i \notin s_1} P(S_1 = s_1)$$

Soit :

$$\boxed{\Pi_i = \Pi_i^1 + E\left(\Pi_i^{2/S_1} 1_{i \notin S_1}\right)}$$

Cette relation met en évidence une dualité d'approche :

- soit on connaît la loi de tirage du 1er échantillon et la loi de tirage du 2ème, conditionnellement au tirage du 1er, et on en déduit par les formules théoriques ci-dessus les probabilités finales d'inclusion des unités.

  Dans la pratique cependant, il est rare que l'on puisse explicitement calculer la probabilité ci-dessus car on ne connaît souvent pas la loi «complète» de S_1 mais seulement ses résumés $? _i^1$.

- soit on connaît les probabilités finales d'inclusion des unités et, connaissant la loi de tirage du 1er échantillon, on cherche la loi de tirage conditionnelle du 2ème qui permette d'obtenir les probabilités finales imposées.

Donnons quelques cas particuliers :

Cas n°1 :

S_1 est de taille fixe n_1, équiprobable sans remise :

$$\Pi_i = \frac{n_1}{N} + \frac{1}{C_N^{n_1}} \sum_{s_1 / i \notin s_1} \Pi_i^{2/s_1} ,$$

avec les sous-cas suivants :

i) conditionnellement à S_1, S_2 est de taille fixe n_2, équiprobable sans remise : $?_i^{2/s_1} = \frac{n_2}{N}$ et $\sum_{s_1 / i \notin s_1} \Pi_i^{2/s_1} = \frac{n_2}{N} C_{N-1}^{n_1}$, d'où :

$$\Pi_i = \frac{n_1}{N} + \frac{n_2}{N} \frac{N - n_1}{N}.$$

ii) idem, mais S_2 est tiré en dehors de S_1 : $\sum_{s_1 / i \notin s_1} \Pi_i^{2/s_1} = \frac{n_2}{N - n_1} C_{N-1}^{n_1}$, d'où :

$\Pi_i = \frac{n_1 + n_2}{N}$ (tout se passe comme si l'on tirait directement un échantillon équiprobable sans remise de taille $n_1 + n_2$).

Cas n° 2 (*proposé par P. ARDILLY*) :

On suppose que $?_i^{2/s_1}$ ne dépend de s_1 que par l'intermédiaire de l'indicatrice $\mathbf{1}_{i \in s_1}$. En d'autres termes, on pose : $?_i^{2/s_1} = ?_i$ si i appartient à s_1

μ_i si i n'appartient pas à s_1

On obtient alors :

$$\Pi_i = \Pi_i^1 + \mu_i \sum_{s_1 / i \notin s_1} P(S_1 = s_1)$$

$$= \Pi_i^1 + \mu_i (1 - \Pi_i^1)$$

ou encore :

$$\boxed{\Pi_i = (1 - \mu_i) \Pi_i^1 + \mu_i}$$

Loi marginale du 2ème échantillon.

Celle-ci sera donnée par :

$$P(S_2 = s_2) = \sum_{s_1} P(S_2 = s_2 \text{ et } S_1 = s_1)$$

$$= \sum_{s_1} P(S_2 = s_2 / S_1 = s_1) P(S_1 = s_1)$$

La probabilité d'inclusion finale d'une unité i dans le 2è échantillon est donnée par :

$$\Pi_i^2 = P(i \in S_2) = \sum_{s_1} P(i \in S_2 \text{ et } S_1 = s_1)$$

$$= \sum_{s_1} P(i \in S_2 / S_1 = s_1)\, P(S_1 = s_1)$$

soit :

$$\Pi_i^2 = \sum_{s_1} \Pi_i^{2/s_1} P(S_1 = s_1)$$

ou :

$$\boxed{\Pi_i^2 = E\Pi_i^{2/S_1}}$$

Comme dans le cas du calcul de Π_i, le calcul explicite de la probabilité ci-dessus est souvent impossible, faute de connaître la loi de S_1. Si l'on reprend les cas particuliers développés ci-dessus où ce calcul est possible, on obtient :

Cas n° 1 : S_1 est de taille fixe n_1, équiprobable sans remise :

$$?_i^2 = \frac{1}{C_N^{n_1}} \sum_{s_1} \Pi_i^{2/s_1} ,$$

i) si S_2 est de taille fixe n_2, équiprobable sans remise, conditionnellement à S_1 :

$$?_i^2 = \frac{1}{C_N^{n_1}} \frac{n_2}{N} C_N^{n_1} = \frac{n_2}{N}.$$

ii) idem, mais S_2 est tiré *en dehors* de S_1 :

$$?_i^2 = \frac{1}{C_N^{n_1}} \frac{n_2}{N - n_1} C_{N-1}^{n_1} = \frac{n_2}{N}.$$

iii) idem, mais S_2 est tiré *dans* S_1 :

$?_i^{2/s_1} = 0$ si $i \notin s_1$

$$= \frac{n_2}{n_1} \text{ sinon,}$$

d'où : $$?_i^2 = \frac{1}{C_N^{n_1}} \frac{n_2}{n_1} C_{N-1}^{n_1-1} = \frac{n_2}{N}.$$

Cas n° 2 : $?_i^{2/s_1} = ?_i$ si i appartient à s_1

μ_i si i n'appartient pas à s_1 :

Alors : $$?_i^2 = ?_i \sum_{s_1 / i \in s_1} P(S_1 = s_1) + \mu_i \sum_{s_1 / i \notin s_1} P(S_1 = s_1)$$

Soit :

$?_i^2 = ?_i\, ?_i^1 + \mu_i (1 - ?_i^1)$.

3. Construction d'estimateurs sans biais.

On cherche à construire un estimateur sans biais du total T, fonction linéaire des observations des deux échantillons. On va donc «réunir » les deux échantillons S_1 et S_2. Dans le cas général, en l'absence de clause spécifique sur les relations entre les deux échantillons, l'échantillon $S_1 \cup S_2$ est potentiellement avec remise car les deux échantillons S_1 et S_2 peuvent ne pas être disjoints.

On peut soit admettre la répétition des unités (en les comptant éventuellement plusieurs fois avec des poids différents), soit ne considérer que des estimateurs sans répétition. Dans ce dernier cas, il faut les construire sous la forme :

$$\hat{T} = \sum_{i \in S_1 \cup S_2} a_i Y_i = \sum_{i \in P} a_i Y_i 1_{i \in S_1 \cup S_2} .$$

3.1 Estimateur de HORVITZ-THOMSON (HT) complet.

En reprenant la logique de l'estimateur de HORVITZ-THOMSON (H-T) sur la réunion des deux échantillons, avec des a_i dont la valeur ne dépend que de l'unité i et pas des réalisations des échantillons S_1 et S_2, la condition d'absence de biais entraînera :

$$a_i = \frac{1}{P(i \in S_1 \cup S_2)} = \frac{1}{\Pi_i} .$$

où ? $_i$ a été calculé au § 2.

3.2 Estimateurs dont les coefficients dépendent de l'appartenance à l'un ou l'autre des deux échantillons.

Une autre logique, qui ne rentre pas dans le cadre du sous-paragraphe précédent, consiste à imposer aux coefficients a_i qui pondèrent les Y_i, de *prendre des valeurs dépendant des réalisations des échantillons S_1 et S_2.*

→ Dans la suite, on va se restreindre à des classes d'estimateurs particuliers en exploitant la situation dans laquelle, l'échantillon S_1 étant tiré initialement et donné une fois pour toutes, le statisticien se trouve confronté ensuite à la nécessité de tirer un 2ème échantillon conditionnellement au premier.

Il paraît donc logique, du fait du caractère « séquentiel » de l'échantillonnage, de supposer que les coefficients des fonctions linéaires choisies comme estimateurs peuvent dépendre de S_1 mais pas de S_2 et qu'ils prennent des valeurs différentes selon l'appartenance de l'unité i à l'un ou l'autre de ces échantillons.

Il faudrait même idéalement distinguer trois cas selon que i appartient à l'un et pas l'autre ou aux deux à la fois.

Un estimateur sans biais du total T satisfaisant aux conditions ci-dessus est donc de la forme :

$$\hat{T} = \sum_{i \in S_1} a_i(S_1) Y_i + \sum_{j \in S_2} b_j(S_1) Y_j$$

$$= \sum_{i\in P} a_i(S_1) Y_i \, 1_{i\in S_1} + \sum_{j\in P} b_j(S_1) Y_j 1_{j\in S_2}$$

$$= \sum_{i\in P} Y_i \left(a_i(S_1) 1_{1_{i\in S_1}} + b_i(S_1) 1_{i\in S_2} \right)$$

On notera que, dans cette formulation, il est possible de compter deux fois une même unité à partir du moment où elle aurait été tirée dans S_1 et dans S_2. Le coefficient qui lui serait alors affecté serait : $a_i\ (S_1) + b_i\ (S_1)$.

On en déduit :

$$E\hat{T} = E\left(E\hat{T}/S_1\right)$$

$$E\left[\hat{T}/S_1 = s_1\right] = \sum_{i\in P} Y_i \left[a_i(s_1) 1_{i\in s_1} + b_i(s_1) \underbrace{E\left(1_{i\in S_2}/S_1 = s_1\right)}_{=\Pi_i^{2/s_1}} \right]$$

D'où :

$$\boxed{E\hat{T} = \sum_{i\in P} Y_i E\left[a_i(S_1) 1_{i\in S_1} + b_i(S_1) \Pi_i^{2/S_1} \right]}$$

L'estimateur $\hat{T}$ est sans biais pour toute variable Y_i si et seulement si :

$$\forall i \in P : \quad \boxed{E\left[a_i(S_1) 1_{i\in S_1} + b_i(S_1) \Pi_i^{2/S_1} \right] = 1} \qquad (1)$$

Cette condition peut aussi s'écrire :

$$E\left[(a_i(S_1) + b_i(S_1) \Pi_i^{2/S_1}) 1_{i\in S_1} + b_i(S_1) \Pi_i^{2/S_1} 1_{i\notin S_1} \right] = 1 \qquad (1 \text{ bis})$$

Sous cette forme, on voit qu'il suffit de déterminer :

$$\begin{cases} a_i(S_1) + b_i(S_1) \Pi_i^{2/S_1} \ \text{pour } i \in S_1 \\ b_i(S_1) \Pi_i^{2/S_1} \ \text{pour } i \notin S_1 \end{cases}$$

On peut encore écrire la relation (1 bis) sous la forme :

$$E\left[a_i(S_1) + b_i(S_1) \Pi_i^{2/S_1} / 1_{i\in S_1} = 1 \right] \ P\{1_{i\in S_1} = 1\}$$
$$+ E\left[b_i(S_1) \Pi_i^{2/S_1} \middle| 1_{i\notin S_1} = 1 \right] \ P\{1_{i\notin S_1} = 1\} = 1$$

Soit :

$$\boxed{E\left[a_i(S_1)+b_i(S_1)\,\Pi_i^{2/S_1}\ /\ i\in S_1\right]\Pi_i^1+E\left[b_i(S_1)\,\Pi_i^{2/S_1}\ /\ i\notin S_1\right]\left(1-\Pi_i^1\right)=1} \qquad (2)$$

A partir de là, toute une gamme de cas particuliers sont envisageables.

<u>1er cas : Estimateur n'utilisant que le 1er sous-échantillon</u> : $b_i(S_1)=0\ \ \forall i$

On obtient alors :

$$E\left(a_i(S_1)1_{i\in S_1}\right)=1\ \ \forall i$$

dont une solution particulière : $a_i(S_1) = a_i$ pour tout i, conduit à : $\boxed{a_i=\frac{1}{\Pi_i^1}}$

On retrouve l'estimateur de HORVITZ-THOMSON relatif au 1er sous-échantillon.

<u>2ème cas : Estimateur n'utilisant que le 2ème sous-échantillon</u> : $a_i(S_1)=0\ \ \forall_i$

Ce cas admet comme sous-cas particulier celui de l'échantillonnage *en deux phases*, où l'on n'utilise au final qu'un échantillon S_2 inclus dans S_1.

On obtient alors :

$$E\left[b_i(S_1)\Pi_i^{2/S_1}\right]=1\ \ \forall i$$

Soit :

$$\sum_{s_1} b_i(s_1)\Pi_i^{2/s_1}\ P(S_1=s_1)\ =1\ \ \forall\, i$$

Ou encore :

$$\sum_{s_1/\,\Pi_i^{2/s_1}\neq 0} b_i(s_1)\Pi_i^{2/s_1}\ P(S_1=s_1)=1\ \ \forall\, i$$

On peut trouver des solutions particulières de plusieurs façons.

- On impose par exemple que, pour chaque réalisation s_1 et pour chaque i tel que $\Pi_i^{2/s_1} \neq 0$, on ait : $b_i(s_1)\ \Pi_i^{2/s_1} = ?_i$ (constante ne dépendant pas de s_1).

  On obtient alors :

  $$?_i \sum_{s_1/\,\Pi_i^{2/s_1}\neq 0} P(S_1=s_1)=1$$

  Soit :

  $$?_i = \frac{1}{P(\Pi_i^{2/S_1}\neq 0)}$$

  On en déduit que :

  $$b_i(s_1) = \frac{1}{\Pi_i^{2/s_1}\ P(\Pi_i^{2/S_1}\neq 0)} \text{ si } \Pi_i^{2/s_1} \neq 0$$

= une valeur indéterminée (par exemple 0) sinon[22].

Au total, on aura :

$$b_i(S_1) = \frac{1_{\Pi_i^{2/S_1} \neq 0}}{\Pi_i^{2/s_1}\, P(\Pi_i^{2/S_1} \neq 0)}$$

Si aucune des variables aléatoires indexées par i, Π_i^{2/S_1}, ne possède de réalisation nulle, on obtiendra plus simplement :

$$b_i(S_1) = \frac{1}{\Pi_i^{2/S_1}} \quad \forall\ i \in \boldsymbol{P}$$

On obtient alors un estimateur de H-T « conditionnel ».

- $b_i(S_1)$ = constante (indépendante de S_1) conduit à :

$$b_i(S_1) = \frac{1}{E(\Pi_i^{2/s_1})} = \frac{1}{\Pi_i^2}.$$

Cet estimateur est en fait l'estimateur de H-T associé au 2ème sous-échantillon, en général peu calculable explicitement (cf. supra).

3ème cas : Estimateur dans lequel le coefficient $a_i(S_1)$ ne dépend pas de S_1 : $a_i(S_1) = a_i \quad \forall i$

On obtient alors :

$a_i\, \Pi_i^1 + E[\, b_i(S_1)\, \Pi_i^{2/S_1}\,] = 1 \quad \forall\ i \in \boldsymbol{P}$

En posant : $a_i\, \Pi_i^1 = a_i$ et $E[\, b_i(S_1)\, \Pi_i^{2/S_1}\,] = 1 - a_i$, avec $a_i \in [0,1]$, on obtient tout d'abord :

$$a_i = \frac{a_i}{\Pi_i^1}.$$

Des solutions particulières pour $b_i(S_1)$ sont obtenues au moyen d'une discussion analogue à celle développée dans le 2ème cas. On obtient ainsi :

-

$$b_i(S_1) = \frac{(1-a_i)1_{\Pi_i^{2/S_1} \neq 0}}{\Pi_i^{2/s_1}\, P(\Pi_i^{2/S_1} \neq 0)}$$

Si aucune des variables aléatoires indexées par i, Π_i^{2/S_1}, ne possède de réalisation nulle, on obtiendra plus simplement :

[22] On notera que, contrairement à l'approche classique de HORVITZ-THOMSON, l'existence de probabilités d'inclusion conditionnelles nulles dans le cadre des échantillonnages « successifs » n'invalide pas la possibilité d'obtenir in fine des estimateurs sans biais. Ce point est important car, dans le cas de la disjonction des deux échantillons, on sera précisément en présence de probabilités d'inclusion conditionnelles nulles.

$$b_i(S_1) = \frac{1-a_i}{\Pi_i^{2/S_1}} \quad \forall\ i \in P$$

- Sous cette dernière hypothèse, la condition $a_i = a \quad \forall i$ fournit un estimateur qui apparaît comme une combinaison barycentrique de l'estimateur de H-T relatif au 1er sous-échantillon et de l'estimateur « conditionnel » obtenu au 2ème cas.

- Le cas $b_i(S_1) = b_i$ conduit à $a_i = \frac{a_i}{\Pi_i^1}$ et $b_i = \frac{1-a_i}{?_i^2}$, en général impossible à calculer explicitement (cf. supra).

4ème cas : Estimateur dans lequel le coefficient $b_i(S_1)$ ne dépend pas de S_1 : $b_i(S_1) = b_i$ pour tout i.

On obtient alors :

$$E\left[a_i(S_1)1_{i\in S_1}\right] + b_i E\,\Pi_i^{2/S_1} = 1$$

soit :

$$E\left[a_i(S_1)1_{i\in S_1}\right] + b_i\,\Pi_i^2 = 1$$

Là encore, les solutions ne sont pas, en général, explicitables, sauf cas particuliers détaillés ci-dessus.

5ème cas : Estimateur dans lequel les coefficients $a_i(S_1)$ et $b_i(S_1)$ ne dépendent pas de S_1 et sont égaux entre eux : $a_i(S_1) = b_i(S_1) = c_i$.

La relation (1) s'écrit alors :

$$\forall\ i \in P : c_i\,[\,E\,(1_{i\in S_1} + \Pi_i^{2/S_1})\,] = 1$$

Soit :

$$c_i\ [\,\Pi_i^1 + E\,(\,\Pi_i^{2/S_1})\,] = 1$$

En remarquant (cf. supra) que : $E\,(\,\Pi_i^{2/S_1}) = \Pi_i^2$, on obtient :

$$c_i = \frac{1}{\Pi_i^1 + \Pi_i^2}$$

On prendra garde ici au fait que les unités peuvent être répliquées, figurant avec le poids c_i au titre du premier échantillon S_1 et à nouveau avec le même poids au titre du second (soit un poids final total de $2\,c_i$). Cependant, dans le cas particulier où les échantillons S_1 et S_2 sont disjoints, il n'y a plus de réplication d'unités, et l'on peut écrire :

$$\Pi_i^1 + \Pi_i^2 = P\,(i \in S_1) + P\,(i \in S_2) = P\,(i \in S_1 \cup S_2) = \Pi_i .$$

On a donc : $c_i = \frac{1}{\Pi_i}$ et l'on retrouve la formule habituelle de l'estimateur de HORVITZ-THOMSON introduite au § 3.1, qui constitue donc bien un cas particulier de l'approche développée dans le présent paragraphe.

3.3 Prise en compte de relations particulières entre les deux échantillons.

Dans tout ce qui précède, on a «réuni » les deux échantillons S_1 et S_2 et on a donné des pondérations spécifiques aux unités selon leur appartenance à l'un ou l'autre, sans tenir compte des relations entre les deux échantillons. En particulier, cette approche englobe les cas où les deux échantillons S_1 et S_2 ne sont pas disjoints. Dans ce cas, une unité i peut être "répliquée", comptant une fois avec la pondération $a_i(S_1)$ et une autre avec la pondération $b_i(S_1)$.

On peut en fait trouver d'autres estimateurs :

- soit en interdisant la réplication des unités, ce qui conduira à imposer des expressions différentes des coefficients selon l'appartenance de l'unité à S_1 et pas à S_2, à S_2 et pas à S_1 ou à $S_1 \cap S_2$
- soit encore en regardant le cas particulier où *le 2ème échantillon est inclus dans le premier (échantillonnage en deux phases)*
- soit en exploitant la situation particulière où *les deux échantillons sont disjoints :* $S_1 \cap S_2 = \emptyset$.

$\rightarrow$ *C'est ce dernier cas que l'on va regarder en détail puisqu'il correspond à la configuration de l'EMEX.*

Dans ce cas, pour toute réalisation de l'échantillon S_1, soit s_1, et pour toute unité i appartenant à s_1, on aura : Π_i^{2/s_1} = 0. On en déduit que, pour tout s_1 et tout i dans ***P*** : $1_{i \in s_1}\ \Pi_i^{2/s_1}$ = 0, soit encore, pour tout i : **la variable aléatoire $1_{i \in S_1}\ \Pi_i^{2/S_1}$ est nulle.**

L'équation (1 bis) va donc s'écrire, pour tout i dans ***P*** :

$$E\left[a_i(S_1)1_{i\in S_1} + b_i(S_1)\,\Pi_i^{2/S_1}1_{i\notin S_1}\right] = 1$$

On peut aussi écrire cette équation sous la forme :

$$E\left[a_i(S_1)/i \in S_1\right]\Pi_i^1 + E\left[b_i(S_1)\Pi_i^{2/S_1}/i \notin S_1\right]\left(1-\Pi_i^1\right) = 1$$

On peut à nouveau trouver des familles de solutions particulières.

1er cas : on impose que les $a_i(S_1)$ ne dépendent de S_1 que par l'intermédiaire de l'indicatrice $\mathbf{1}_{i \in s_1}$, c'est-à-dire peuvent s'écrire : $a_i(S_1) = a_i 1_{i \in S_1}$. En d'autres termes, pour tous les i appartenant S_1, la valeur $a_i(S_1)$ dépend de i mais plus de S_1.

L'équation ci-dessus devient alors, pour tout i dans ***P*** :

$a_i\ \Pi_i^1$ + E [$b_i(S_1)\ \Pi_i^{2/S_1}\ 1_{i \notin S_1}$] = 1.

En posant : $a_i\ \Pi_i^1 = a_i$ et E [$b_i(S_1)\ \Pi_i^{2/S_1}\ 1_{i \notin S_1}$] = 1 - a_i , avec $a_i \in$ [0,1], on obtient tout d'abord :

$a_i = \frac{a_i}{\Pi_i^1}$.

Puis on traite le 2ème terme comme on l'a fait dans les cas 2 et 3 du § 3.2 :

$$E[\, b_i(S_1)\ \Pi_i^{2/S_1}\ 1_{i \notin S_1}\,] = \sum_{s_1} b_i(s_1)\Pi_i^{2/s_1} 1_{i \notin s_1} P(S_1 = s_1)$$

$$= \sum_{s_1 / i \notin s_1} b_i(s_1)\Pi_i^{2/s_1}\ P(S_1 = s_1)$$

$$= \sum_{\substack{i \notin s_1 \\ s_1 / \Pi_i^{2/s_1} \neq 0}} b_i(s_1)\Pi_i^{2/s_1}\ P(S_1 = s_1)$$

- On impose par exemple que, pour chaque réalisation s_1 et pour chaque i tel que :

  $i \notin s_1$ et $\Pi_i^{2/s_1} \neq 0$, on ait : $b_i(s_1)\ \Pi_i^{2/s_1} = ?_i$ (constante ne dépendant pas de s_1).

  On obtient alors :

  $$?_i \sum_{\substack{i \notin s_1 \\ s_1 / \Pi_i^{2/s_1} \neq 0}} P(S_1 = s_1) = 1 - a_i .$$

  On obtient alors :

  $$?_i = \frac{1 - a_i}{P(i \notin S_1\ et\ \Pi_i^{2/S_1} \neq 0)}$$

  D'où :

  $$b_i(s_1) = \frac{1 - a_i}{\Pi_i^{2/s_1} P(i \notin S_1\ et\ \Pi_i^{2/S_1} \neq 0)} \text{ si } \Pi_i^{2/s_1} \neq 0$$

  = une valeur indéterminée (par exemple 0) sinon.

  Au total, on aura :

$$b_i(S_1) = \frac{(1 - a_i) 1_{\Pi_i^{2/S_1} \neq 0}}{\Pi_i^{2/S_1} P(i \notin S_1\ et\ \Pi_i^{2/S_1} \neq 0)}$$

On note en particulier que : $b_i(S_1) = 0$ dès que $\Pi_i^{2/S_1} = 0$, ce qui est obligatoirement le cas pour tous les i appartenant à S_1.

Si, pour toute réalisation s_1 et pour tout i n'appartenant pas à s_1, on a : $\Pi_i^{2/s_1} \neq 0$, alors cette valeur se simplifie et l'on obtient :

$$b_i(S_1) = \frac{(1 - a_i) 1_{i \notin S_1}}{\Pi_i^{2/S_1} P(i \notin S_1)}$$

Soit :

$$b_i(S_1) = \frac{(1-a_i)1_{i\notin S_1}}{\Pi_i^{2/S_1}(1-\Pi_i^1)}$$

Sous cette dernière hypothèse, on peut expliciter encore deux cas particuliers :

- la condition $a_i = a \ \forall i$ fournit l'estimateur

$$\hat{T} = a\sum_{i\in S_1}\frac{Y_i}{\Pi_i^1} + (1-a)\sum_{i\in S_2}\frac{Y_i}{\Pi_i^{2/S_1}\left(1-\Pi_i^1\right)}$$

qui apparaît comme une combinaison barycentrique de l'estimateur de H-T relatif au 1er échantillon et d'un deuxième estimateur, également de type H-T, relatif au 2ème échantillon. Les cas limites a = 0 ou a = 1 sont évidemment licites.

$$a = 0 \ : \ \hat{T} = \sum_{i\in S_2}\frac{Y_i}{\Pi_i^{2/S_1}\left(1-\Pi_i^1\right)}$$

$$a = 1 \ : \ \hat{T} = \sum_{i\in S_1}\frac{Y_i}{\Pi_i^1}$$

(solution analogue à celle du 1er cas du § 3.2)

- la condition $a_i = ?_i^1 \ \forall i$ fournit l'estimateur :

$$\hat{T} = \sum_{i\in S_1} Y_i + \sum_{i\in S2}\frac{Y_i}{\Pi_i^{2/S_1}}$$

Tout se passe comme si S_1 était une population de référence : on prend le total non pondéré sur cette sous-population et on rajoute une estimation de type H-T conditionnel relatif au 2ème échantillon.

2ème cas :

On conserve la même hypothèse que précédemment sur les $a_i(S_1)$, d'où les valeurs :

$a_i(S_1) = \frac{a_i}{\Pi_i^1}$, et l'on impose que $b_i(S_1)$ prenne également une valeur b_i indépendante de S_1 lorsque $i \notin S_1$. On trouve alors :

$$b_i(S_1) = \frac{(1-a_i)1_{i\notin S_1}}{E(\Pi_i^{2/S_1}1_{i\notin S_1})}$$

On a vu dans le § 1 que le dénominateur de cette expression valait : Π_i - Π_i^1, soit : Π_i^2 (puisque les deux échantillons S_1 et S_2 sont disjoints).

On obtient finalement l'estimateur :

$$\hat{T} = \sum_{i\in S_1}\frac{a_i}{?_i^1}Y_i + \sum_{i\in S_2}\frac{1-a_i}{?_i^2}Y_i$$

dont le cas particulier $a_i = a$ conduit à une combinaison barycentrique de deux estimateurs de H-T. calculés séparément sur les deux échantillons, qui apparaissaient comme limites du 2ème cas du § 3.2. On se heurte néanmoins, comme déjà indiqué, à l'impossibilité de calculer en général $?_i^2$.

ANNEXE 4 : conservation de la propriété d'équilibrage dans le cas d'un tirage équilibré conditionnel et à probabilités égales.

Résultat :

Si l'on tire un échantillon s_1 dans une population ***P*** selon un tirage équilibré sur une variable X et que le tirage soit à probabilités égales ;

Si l'on tire un deuxième échantillon (s_2) dans ***P*** \ s_1 tel que le tirage conditionnel de s_2 sachant s_1 soit :

- à probabilités égales
- équilibré sur la même variable X que l'échantillon s_1 (et au niveau de ***P*** \ s_1) ;

Alors $s = s_1 \cup s_2$ est équilibré sur ***P*** (sous la loi de tirage finale de s) pour la variable X.

Démonstration :

Notons Π_i^1 la probabilité d'inclusion d'une unité quelconque i dans s_1, Π_i^{2/s_1} celle de l'unité i dans s_2 (conditionnellement au tirage de s_1) et enfin Π_i, la probabilité finale d'inclusion relative à $s = s_1 \cup s_2$.

On veut montrer que :

$$\sum_{i \in s_1 \cup s_2} \frac{x_i}{\Pi_i} = \sum_{i \in P} x_i$$

Du fait des modes de tirage, les différentes probabilités ne dépendent plus de l'unité i à laquelle elles se réfèrent : on notera donc : $\Pi_i = \Pi$, $\Pi_i^1 = \Pi^1$, Π_i^{2/s_1} Π^{2/s_1}.

On a montré, lorsque les deux tirages successifs sont de taille fixe, équiprobables sans remise, que :

$\Pi = \Pi^1 + \Pi^{2/s_1} (1 - \Pi^1)$.

(cf. Annexe 3, § 2, cas n°1, ii)).

La propriété d'équilibrage sur s_1 s'écrit :

$$\frac{1}{\Pi^1} \sum_{i \in s_1} x_i = \sum_{i \in P} x_i \quad (1)$$

ou, de façon équivalente :

$$\sum_{i \in P - s_1} x_i = (1 - \Pi^1) \sum_{i \in P} x_i \quad (2).$$

La propriété d'équilibrage sur s_2 s'écrit :

$$\frac{1}{\Pi^{2/s_1}} \sum_{i \in s_2} x_i = \sum_{i \in P - s_1} x_i$$

ou, de façon équivalente en utilisant (2) puis (1) :

$$\sum_{i \in s_2} x_i = \Pi^{2/s_1} (1 - \Pi^1) \sum_{i \in P} x_i$$

d'où :

$$\sum_{i \in s_1} x_i + \sum_{i \in s_2} x_i = \Pi^1 \sum_{i \in P} x_i + \Pi^{2/s_1} (1 - \Pi^1) \sum_{i \in P} x_i$$

$$= [\Pi^1 + \Pi^{2/s_1} (1 - \Pi^1)] \sum_{i \in P} x_i$$

soit, en vertu de la disjonction des deux échantillons s_1 et s_2 :

$$\boxed{\sum_{i \in s_1 \cup s_2} x_i = \Pi \sum_{i \in P} x_i .}$$

CQFD.

ANNEXE 5 : Eléments de volumétrie

des deux bases EM et EMEX.

SGESTION = 0

	Nombre d'UP	Nombre de COM	Nombre de logements recensés	Nombre de logements recensés après disjonction Emploi
EM	128	2018	451 776	**447 958**
EMEX	253	3557	877 624	**866 699**
EM+EMEX	381	5575	1 329 400	**1 314 657**

SGESTION = 1

	Nombre d'UP	Nombre de COM	Nombre de logements recensés	Nombre de logements recensés après disjonction Emploi
EM	75	284	444 978	**439 671**
EMEX	151	542	866 118	**853 635**
EM+EMEX	226	826	1 311 096	**1 293 306**

SGESTION = 2

	Nombre d'UP	Nombre de COM	Nombre de districts//groupes	Nombre de logements recensés	Nombre logemer recensés a disjoncti Emplo
EM	93	610	7 839 // 2 790	497 962	**490 211**
EMEX	87	487	7 047 // 2 607	486 943	**479 050**
EM+EMEX	180 (exhaustif)	1097	14 886 // 5 397	984905	**969 261**

NB : Nombre de districts dans la base de tirage (après disjonction avec districts EM) = 23 077, soit 8 118 groupes de districts.

SGESTION = 3

	Nombre d'UP	Nombre de COM	Nombre de districts	Nombre de logements recensés	Nombre logemer recensés a disjonction I
EM	52		5 098	399 845	**394 558**
EMEX	52		5 081	399 486	**394 420**
EM+EMEX	52		10 179	799 331	**788 978**

NB : Nombre de districts dans la base de tirage (après disjonction avec districts EM) = 96 821.

SGESTION = 4

	Nombre d'UP	Nombre de COM	Nombre de districts	Nombre de logements recensés	Nombre logemen recensés a disjoncti Emploi
EM	1		2 465	228 328	**225 818**
EMEX	1		2 472	228 440	**226 698**
EM+EMEX	1		5 937	456 768	**452 516**

NB : Nombre de districts dans la base de tirage (après disjonction avec districts EM) = 46 811.