LES ENQUETES PAR PANEL : EN QUOI DIFFERENT-ELLES DES AUTRES ENQUETES ? suivi de : comment attraper une population en se servant d'une autre

Jean-Claude Deville

Le terme "*panel*" est souvent utilisé de façon abusive. Dans certains milieux un peu snob de la pub, il est utilisé pour échantillon. En économétrie, dans la locution "données de panel", il se réfère à des modèles à deux indices dont l'un est ordonné. On appréciera au passage l'usage délicieux du mot "données" : données par qui ? comment ? Le but de cet exposé est de montrer les contraintes logiques liées à l'élaboration de ces fameuses "données".

Un minimum de clarifications s'impose. Après quelques constatations de bon sens sur la notion de temps et d'identification, on essaiera de passer en revue les différentes formes de collecte où le temps intervient : enquêtes répétées, enquêtes de cohortes, échantillons rotatifs ou coordonnés, enquêtes continues, panels.

On introduira un formalisme unificateur de toutes ces formes d'enquête. Vu sous un certain angle, les panels apparaissent comme des enquêtes ordinaires justiciables des méthodes les plus habituelles d'analyse, d'estimation, d'estimation de variance et même d'échantillonnage.

Les exemples seront pris de préférence dans le domaine des enquêtes auprès des personnes, mais sont transposables aux enquêtes auprès des autres agents économiques. On verra en particulier pourquoi la notion de panel de ménage est à peu près vide de sens, bien qu'on puisse assez facilement faire des statistiques de ménages à l'aide d'un échantillon de personnes.

1 - Population et identifiant

On appelle généralement panel (cf. [LAVALLEE (1996)]) toute enquête où les unités sont enquêtées à plusieurs dates successives. Cette définition, qui ne prend en compte que le processus de collecte des données, reste tout à fait imprécise si nous voulons examiner les problèmes statistiques liés à l'introduction du temps, c'est-à-dire de variables datées, dans les enquêtes par sondage.

Il faut donc retourner à la racine des choses. La période d'étude T est une intervalle de temps $T = (t_0, t_f)$ où t_f est généralement fini (c'est l'horizon de l'étude) mais souvent, en pratique, indéfini et donc pris égal à l'infini (ce qui ne manque pas de poser des problèmes angoissants, comme chaque fois qu'on est, dans la pratique, confronté à ce mystère). A chaque instant t est associée une population finie identifiée U_t, c'est-à-dire l'objet potentiel d'un sondage. A chaque individu k (k est l'"étiquette", le "label" ou l'identifiant) de U_t est associé un vecteur de variables d'intérêt $y_k^{(t)}$, un vecteur de variables auxiliaires $x_k^{(t)}$; éventuellement on dispose aussi d'information auxiliaire externe $Z^{(t)}$.

Un système d'enquêtes répétées consiste à réaliser à diverses dates $t \in F = \{t_1, t_2, \ldots, t_n, \ldots\}$ (où F est fini (dans tout intervalle fini de T dans le cas où $t_f = \infty$!)), des enquêtes par sondage indépendantes les unes des autres dans les $U_t (t \in F)$ destinées à recueillir les "mêmes" variables. C'est la pratique traditionnelle des recensements successifs de la population (au ¼ !) ou des enquêtes-logements. Le but de l'exploitation statistique est alors de fixer des niveaux de y aux différentes périodes et de les comparer pour évaluer des évolutions globales. Dans le cas où certaines unités statistiques auxiliaires sont pérennes (zone d'emploi, communes), le cumul de plusieurs enquêtes successives peut aider à élaborer des statistiques "locales". Il n'y a pas grand chose à dire de plus sur cette façon de traiter le temps, sinon qu'il ne joue aucun rôle particulier : on pourrait dire que t désigne un pays européen par exemple. Autrement dit on ne suppose absolument rien - dans la technique d'enquête en tout cas - sur le lien entre les populations $U_t (t \in F)$.

La spécificité du temps (implicite mais qu'il vaut mieux expliciter) est que la population U_t varie en un certain sens "continûment". Soyons un peu formel : soit $U = \bigcup_{t \in T} U_t$ la population d'étude. On dira que la population est "renouvelée" si U est fini (ou, si T est non borné, si $U_B = \bigcup_{t \in B} U_t$ est fini pour toute partie bornée B de T, (on a déjà des problèmes avec l'infini !)), autrement dit si une étiquette k de U figure dans une partie T_k assez grosse. On ne restreindra guère la généralité en supposant que T_k est un intervalle (ou, éventuellement, une réunion finie d'intervalles *(Exemple : Etudes sur les chômeurs)*).

Ceci suppose implicitement qu'il existe un moyen d'identifier les individus de U, les unités statistiques, comme étant les mêmes à deux époques distinctes et donc au cours de toute la période d'étude. Cette condition, théorique et pratique, est essentielle, fondatrice, dès qu'on veut aller au delà des études par enquêtes répétées et qu'on veut travailler sur des données dites longitudinales.

Cette remarque a des conséquences immédiates sur la façon d'envisager la statistique longitudinale des populations $\neq$ humaines. Il paraît à peu près clair qu'on puisse identifier d'une façon rigoureuse les personnes. Sans aller jusqu'à des arguments génétiques, l'usage du NIR ou simplement du "nom, prénom, date et lieu de naissance" semble faire l'affaire. (Ceci dit, on peut se poser des questions sur des époques ou des pays où on ignore les tests génétiques et où les inscriptions d'Etat-Civil sont un tantinet déficientes !). Encore faut-il pouvoir, techniquement, utiliser cet identifiant. Ce n'est par exemple guère possible de façon massive dans les recensements ; ceux-ci autorisent néanmoins la confection de l'Echantillon Démographique Permanent (E.D.P, "Le Panel Démographique" pour les initiés").

L'identification des ménages pose un tout autre problème qui n'a vraisemblablement pas de solution exempte d'arbitraire. Stricto sensu, un ménage est un ensemble de personnes et s'identifie par la liste de ses membres. Un ménage est donc le même à t et t' s'il est composé des mêmes personnes.

Cette définition a l'avantage de prendre en compte les déménagements. En revanche toute arrivée ou tout départ signifie la "mort" du ménage et la "naissance" d'un autre, même en cas de "mouvement naturel" au sens démographique habituel : "naissances" ou "décès" c'est-à-dire entrées et sorties de la population d'étude (ces notions peuvent se formaliser facilement de façon rigoureuse, mais nous n'insisterons pas ici là-dessus).

Le "solde naturel" de la population entre t et $t'\left(t<t'\right)$ est la différence symétrique $S_{t,t'}=U_t \Delta U_{t'}=\left\{k;\left(k\in U_t \text{ et } k\notin U_{t'}\right)\right.$ (décès) ou $\left.\left(k\notin U_t \text{ et } k\in U_{t'}\right)(\text{naissances})\right\}$. On peut dire qu'un ménage reste identique à lui-même, et donc peut conserver le même identifiant, s'il n'est affecté que par le mouvement naturel. Cette notion n'est pas si simple à formaliser, et donc à vérifier concrètement dans une opération d'enquête : elle est en effet différentielle en ce sens qu'elle demande la prise en compte de tous les "événements démographiques" survenus ente t et t' et mettant en cause les individus susceptibles d'avoir appartenu à ce ménage longitudinal au cours de la période. Identifier les ménages m_t et $m_{t'}$ lors de deux recensements, par exemple, est impossible même si la population $S_{t,t'}$ est parfaitement connue ! La seule définition possible devrait contenir la clause : $m_t \Delta m_{t'} \subset S_{t,t'}$. Mais si $m_t \cap m_{t'}=\emptyset$ on ne sait pas si $m_{t'}$ comporte les personnes "nées du ménage m_t", celles qui y appartenaient à t étant décédées, ou si il s'agit d'un remplacement complet (les problèmes concret de ce type concernent les phénomènes de migrations plus que les naissances-décès habituels). On peut vérifier que la définition suivante :

$$\exists\, t_o=t<t_1<t_2 \quad <t_n=t' : \forall\, i\,(i=0,\ldots,n-1) : m_{t_i} \Delta\, m_{t_{i+1}} \subset S_{t_i,\, t_{i+1}} \text{ et } m_{t_i} \Delta\, m_{t_{i+1}} \neq \emptyset \ ,$$

capture correctement la notion de mouvement naturel. Reste à formaliser les autres "événements démographiques" que peuvent subir les ménages et les critères d'identification logique qu'on peut mettre en œuvre pour établir une filiation des identifiants. Ce n'est pas de la tarte. Prenons pour exemple un cas simple : l'individu $k \in m_{t-\varepsilon}$ se met en ménage avec l'individu $\ell \in m'_{t-\varepsilon}$. C'est la seule modification qui touche la population $m_{t-\varepsilon} \cup m'_{t-\varepsilon}$, de sorte qu'à $t+\varepsilon$ on a trois ménages :

$$m^1_{t+\varepsilon} = m_{t-\varepsilon} - \{k\}$$
$$m^2_{t+\varepsilon} = m_{t-\varepsilon} - \{\ell\}$$
$$m^3_{t+\varepsilon} = \{k, \ell\}.$$

On voit facilement que l'identification de ce qu'on peut appeler des "nouveaux ménages" est très arbitraire. On remarque par exemple $m^1_{t+\varepsilon}$ et $m^2_{t+\varepsilon}$ peuvent très bien être vides . Le concept de "chef de ménage" ou de "personne de référence " peut aider à ces problèmes d'identification, mais on sait bien à quel point leurs définitions sont arbitraires.

L'identification de logements est relativement plus facile. Ils sont généralement repérés par une adresse, un identifiant fiscal ou tout autre. Les notions de construction, d'achèvement et de destruction sont assez rigoureuses et correspondent au "mouvement naturel" de la population. Les seuls problèmes (hormis celui des habitations mobiles peut-être ?) est celui des regroupements et éclatements de logements. Ces événements sont cependant suffisamment rares et faciles à traiter statistiquement (voir § 8) pour ne pas gêner l'analyse longitudinale.

De façon générale, l'analyse longitudinale suppose une codification rigoureuse des événements démographiques survenant à la population d'études. On peut alors de façon non ambiguë et cohérente établir des règles de création, de suppression et de filiation des identifiants c'est-à-dire des unités statistiques de la population d'étude.

C'est possible pour les personnes et, dans une large mesure, pour les logements. On est très loin de ce pré-requis en ce qui concerne les ménages de sorte que c'est au moins un abus de langage (et au plus un non-sens) que de parler de "panel de ménages".

2 - Population fixe dans le temps

La population d'étude est constante (dans le temps) si $U_t = U_{t_0} = U_{t_f}$ pour tout t de T. On parle alors classiquement d'une cohorte. Cette situation résulte très généralement d'une convention qui définit cette cohorte à partir d'une population évolutive : ensemble des personnes qui ont subi le même événement au même

moment (c'est la définition classique des démographes), personnes survivantes à la date t_r , personnes figurant dans la même liste à la date t_0 (les "conscrits"), voire même étoiles d'un catalogue pour un panel à visées astronomiques.

Pour ce type de population, les problèmes sont relativement simplifiés. L'échantillonnage ne pose aucun problème particulier puisqu'on peut en principe travailler sur une base de sondage fixe et connue. Le recueil d'information, en revanche, n'a pas de solution simple et peut revêtir plusieurs aspects.

La méthode la plus rapide et la moins coûteuse est celle de l'enquête rétrospective. Son plus beau fleuron à l'Insee est sans doute, dans sa forme classique, l'enquête sur les familles associée aux recensements pour l'étude de la fécondité, de la nuptialité et de l'activité des femmes. Le principal problème qu'elle pose est celui des erreurs de mémoire (le "biais" de mortalité étant négligeable, semble-t-il). On pourra se référer par exemple à [J.C Deville (1972)].

Dans les méthodes d'observation suivie on peut distinguer les techniques d'observation longitudinales, qui sont des cas particulier de panels, comme les panels d'élèves suivis par les services de l'Education Nationale. Elles se caractérisent par le fait que l'échantillon est unique pour toute la période d'étude et que des données sont collectées pour l'ensemble des dates d'observation.

Les autres méthodes peuvent être qualifiées d'observation partielle. On y observe les données relatives à une unité échantillonnée k uniquement à certaines dates dépendant de k. Les méthodes les plus fréquemment utilisées sont celle des échantillons tournants (ou rotatifs) et celle des échantillons coordonnés, qui est plus générale.

Un échantillon rotatif résulte du partage d'un échantillon global s (exemple : les aires de l'enquête-emploi) en sous-échantillons s_a suivis sur une sous-période fixée T_a de T. Chaque s_a est en quelque sorte un panel sur la période T_a et l'échantillon rotatif peut être vu comme une famille de panels qui se chevauchent, c'est-à-dire relatifs à des périodes T_a qui se recouvrent. L'échantillon rotatif est conçu pour être extrapolable à chaque période d'observation $F \subset T$ et posséder les mêmes propriétés statistiques. Ceci implique (en première approximation et sous certaines conditions) que les sous-échantillons soient de même taille et observés le même nombre de fois.

Les schémas de rotation peuvent aller du très simple (enquête emploi annuelle classique) à l'assez compliqué (future enquête emploi en continu) si on veut obtenir une efficacité dans la mesure des évolutions à court terme et à moyen terme (trimestriel, mensuel ou annuel).

Dans un échantillonnage coordonné dans le temps, la période d'observation de chaque unité k est définie indépendamment de toute référence à l'appartenance à un sous échantillon. La méthode des numéros aléatoires est fréquemment utilisée dans les enquêtes auprès des entreprises et permet, par exemple, de faire entrer en observation une entreprise qui grossit en fonction de la taille qu'elle atteint. On se reportera à [Cotton, Hesse (1992)] et à [Hesse (1994)] pour plus de détails.

3 - Enquêtes longitudinales et enquêtes continues

Ça n'a quasiment rien à voir.

Dans tout ce papier (sauf ce paragraphe) nous supposons qu'il n'y a aucune ambiguïté sur la définition du temps et de la date d'observation. En particulier chaque vague d'enquête a lieu à une date bien déterminée $t \in F \subset T$. Quand on y regarde d'un peu plus près, ça change. Les différentes unités ne sont pas enquêtées au même moment exactement. On s'en tire en admettant que les variations possibles de date d'enquête n'induisent que des variations minimes sur les variables d'intérêt. On tolère donc une erreur de mesure $\left(\text{collecter } y_k^{t+\varepsilon} \text{ au lieu de } y_k^{(t)}\right)$ qu'on juge négligeable. On peut aussi chercher, par appel à la mémoire, à obtenir à la date $t + \varepsilon$ la valeur de $y_k^{(t)}$. Ceci revient à tolérer une autre erreur de mesure liée au biais de mémoire et à estimer qu'elle est négligeable.

Dans certains cas, en fait quand des variations rapides des variables d'étude doivent être prises en compte, cette tolérance devient dangereuse. On sait par exemple, que la date de collecte de la défunte enquête-emploi trimestrielle est une des causes de l'instabilité des résultats qu'elle faisait apparaître. L'existence de phénomènes saisonniers instables (liés au climat : "la" vague de froid annuelle, la date des vendanges pour prendre un exemple célèbre [Le Roy-Ladurie (1967)]) peut perturber complètement la statistique. En particulier une collecte à date fixe peut rater complètement un phénomène saisonnier survenant plutôt tard ou, au contraire, de façon exceptionnellement précoce. Une façon de ne pas rater ces variations temporelles de survenue d'événements inévitables est de réaliser une opération d'enquête à collecte continue.

L'exemple le plus classique (du moins théoriquement car la réalisation pratique laisse beaucoup à désirer) est celui des enquêtes sur les budgets de famille. Dans ces enquêtes, l'échantillon est réparti de façon plus ou moins aléatoire sur la période d'étude. On observe donc $y_k^{(t_k)}$ où (k, t_k) est un aléatoire fixé par le plan de sondage. Le but est d'estimer des quantités de la forme $\sum_U \int_T y_k^{(t)}\, dt$ à partir des observations. Le caractère aléatoire des t_k fixés par le plan permet d'obtenir des

estimateurs sans biais à partir du moment où toute date t a, en un certain sens, une probabilité non nulle de faire l'objet d'une enquête.

Le fait de réaliser une enquête en continu n'exclut pas celui de recueillir des données longitudinales de type panel. C'est en principe, ce qui sera fait pour la future enquête-emploi. Et ce qui fait aussi une des difficultés, complètement sous-estimée, de cette opération.

4 - Populations renouvelées dans le temps et panels

Un panel est un échantillon adapté à l'étude d'une population qui se renouvelle au cours de la période d'étude T. En particulier il a la propriété d'être extrapolable à toute population U_t pour $t \in F$ (date d'observation) - *propriété transversale* - et de contenir toutes les données relatives à une unité k pour toutes dates d'observation $F_k = \{t; t \in F \text{ et } k \in \bigcup_t\} = F \cap T_k$ où l'unité k appartient à la population d'étude - *propriété longitudinale* - .

Ces exigences semblent a priori beaucoup plus fortes que tout ce que nous avons envisagé jusqu'à maintenant. C'est tout à fait vrai en ce qui concerne la pratique de l'échantillonnage, de la collecte et de la gestion des données (y compris leur analyse). Sur le plan théorique, une petite astuce rend les choses beaucoup plus manipulables et permet de définir un cadre logique où tout s'organise assez bien.

Poursuivons la discussion commencée au paragraphe [1]. La période d'étude T est fixée, la population d'étude $U = \bigcup_{t \in T} U_t$ aussi. Un plan de sondage de type panel est donc une loi de probabilité p(s) sur l'ensemble des parties de U. Celle-ci permet de définir les probabilités d'inclusion π_k des individus dans le panel, ainsi que les probabilités d'inclusions doubles $\pi_{k\ell}$ permettant en principe de se livrer à des calculs et à des estimations de variance.

A chaque individu k de U est associée l'indicatrice $T_k = \{t \in T : k \in U_t\}$. C'est , comme on l'a dit, un intervalle ou une réunion d'intervalles. L'indicatrice d'inclusion est la famille de variables $E_k^{(t)} = 1$ si $k \in U_t$ et zéro sinon. Elle nous livre directement les statistiques à établir sur l'état de la population. L'effectif de celle-ci à l'époque t est $N_t = \sum_U E_k^{(t)}$ qui est estimé par $\hat{N}_t = \sum_U \frac{E_k^{(t)}}{\pi_k}$ si on utilise l'estimateur de Horvitz-Thompson. Les variables d'intérêts (ou auxiliaires) $y_k^{(t)}$ sont définies et collectables uniquement sur T_k, en principe. Complétons les de façon arbitraire sur $T - T_k$. On obtient ainsi une structure identique à celle des populations fixes dans le

temps. Dans notre formalisme celles-ci se caractériseraient par le fait que pour tout k, $E_k^{(t)}$ est constante et égale à 1. Une caractéristique fixe de k est une variable $x_k^{(t)}$ constante sur T_k (et sur T). Exemple : Lieu de naissance.

On peut dire aussi que $\left(U, y_k^{(t)}\right)$ est un processus aléatoire au sens mathématique du terme (la tribu sur U étant l'ensemble des parties). D'une certaine manière, contrairement au célèbre article de [G. KALTON (1993)], nous avons procédé à la suppression de la quatrième dimension.

L'intérêt de cette construction est le suivant. On sait que toutes les statistiques qui présentent un certain intérêt dans les problèmes de sondage peuvent être vues comme des totaux ou des fonctions de totaux. C'est le cas des effectifs, des moyennes, des ratios, des taux d'évolution, des corrélations entre variables (et donc des corrélations temporelles). Dans le cas des populations évolutives, on ne s'intéressera (sauf peut-être exception - le concours est ouvert) qu'à des "totaux vivants" c'est-à-dire de la forme $\sum_U y_k^{(t)} E_k^{(t)}$. Bien évidemment tous ces totaux sont estimables dès que l'estimateur de Horvitz-Thompson est disponible, c'est-à-dire que les π_k sont connus.

Donnons juste quelques exemples.

Le total de y à t est $\sum_U y_k^{(t)} E_k^{(t)} = \sum_U y_k^{(t)} = Y_t$. Il s'estime par $\hat{Y}_t = \sum_s y_k^{(t)} E_k^{(t)} / \pi_k = \sum_{s_t} \frac{y_k^{(t)}}{\pi_k}$ où $s_t = s \cap U_t$. On remarque que π_k ne dépend pas de t ce qui est la caractéristique d'un panel (en fait certaines variables auxiliaires, éventuellement datées, présentes dans la base de sondage peuvent déterminer le choix des π_k, comme dans tout problème d'échantillonnage. Nous ne développerons pas plus ces considérations dans ce papier). La moyenne des $y_k^{(t)}$ est le ratio $\frac{Y_t}{N_t} = \overline{Y}_t$ et s'estime par $\hat{\overline{Y}}_t = \frac{\hat{Y}_t}{\hat{N}_t}$.

L'évolution nette de la moyenne de y entre t et s est $\overline{Y}_t - \overline{Y}_s$ et s'estime par $\hat{\overline{Y}}_t - \hat{\overline{Y}}_s$. L'évolution brute est l'évolution de la moyenne entre t et s de la population présente aux deux époques.
C'est le ratio $\frac{\Delta_{ts} Y}{N_{ts}}$ où $N_{ts} = \sum_U E_k^{(t)} E_k^{(s)}$ et $\Delta_{ts} Y = \sum_U \left(y_k^{(t)} - y_k^{(s)}\right) E_k^{(t)} E_k^{(s)}$ (alors que pour l'accroissement de y c'est $Y^t - Y^s = \sum_U \left(y_k^{(t)} E_k^{(t)} - y_k^{(s)} E_k^{(s)}\right)$). Toutes ces

quantités sont des totaux ou des fonctions de totaux et s'estiment donc de façon naturelle et simple à partir de l'estimateur de Horvitz-Thompson. Le calcul et l'estimation de la variance de ces statistiques ne pose rigoureusement aucun problème nouveau.

Bref, tant qu'on se limite à l'estimateur de Horvitz-Thompson et que les données sont complètes (éventuellement imputées), on n'a aucun problème particulier avec les données de panel.

5 - Modalités d'exploitation d'un panel, nature et rôle de l'information auxiliaire et correction de la non-réponse

On a coutume de poser les deux assertions suivantes :

– un panel doit pouvoir être exploité transversalement,
– un panel doit s'exploiter à sa date d'échéance (ce qui pose problème si $t_r = \infty$!).

Ne faisons pas dans la dentelle : un panel doit pouvoir s'exploiter pour toute famille de date $F_0 \subset F$ à condition que $\sup_F t \leq \theta$ ou θ désigne la date actuelle (moins délai de préparation des données).

Les cas les plus habituels sont :

- $F_0 = \{t^*\}$: où $t^* = \sup \{t \in F ; t \leq \theta\}$: exploitation transversale.
- $F_0 = \{t_0, t^*\}$: Evolution depuis l'instant origine. C'est le cas des indices de type classique, Paasche ou Laspeyres.
- $F_0 = \{t^{**}, t^*\}$ ou $t^{**} = \sup \{t \in F ; t < t^*\}$: évolution récente.
- $F_0 = \{t \in F ; t \leq \theta\}$ ensemble des données connues, analyse longitudinale "complète" à la date actuelle.

Tant que les données sont complètes (non-réponse compensée par une imputation définitive) et qu'on utilise l'estimateur de Horvitz-Thompson, il n'y a aucun problème : toutes les statistiques sont estimables, l'estimation de variance peut se faire (au bémol des imputations près) et, surtout, il y a cohérence parfaite entre toutes les exploitations qu'on peut envisager.

L'ennui, c'est qu'on ne procédera jamais de cette manière.

En effet, le mode de correction de la non-réponse et l'amélioration de l'estimation par incorporation d'information auxiliaire va dépendre du choix de F_0 et de θ (car l'information externe disponible dépend de ce paramètre. On peut par exemple différer une exploitation transversale pour profiter de la disponibilité attendue d'une information auxiliaire - la pyramide des âges à t^* par exemple). Si on ne se pose pas de problèmes de cohérence des statistiques (et donc d'éventuelles révisions des résultats d'exploitations antérieures du panel), nous sommes ramenés à un problème standard : estimer des statistiques portant sur les variables $y^{(t)}$ pour $t \in F_0$ en utilisant, pour la correction des non-réponses et l'estimation, l'information auxiliaire disponible à savoir :

- *Les* $z^{(t)}$ pour les t disponibles $(\leq \theta!)$
- *Les* $x_k^{(t)}$ pour $t \in F_0$
- *Les* $x_k^{(t)}$ *et les* $y_k^{(t)}$ pour $t \in F - F_0$ $(et\ t \leq \theta)$

En particulier les $y_k^{(t)}$ $t \notin F_0$, par exemple les valeurs antérieures de $y_k^{(t)}$ pour une analyse transversale, sont à considérer comme des variables auxiliaires utilisables dans la procédure d'estimation.

Examinons par exemple le cas de l'analyse transversale comparé à celui de l'analyse longitudinale complète à date actuelle.

On peut, conformément aux habitudes, décider de compenser la non-réponse complète par repondération. Celle-ci utilisera les données auxiliaires disponibles pour l'époque t^* mais éventuellement aussi un modèle de réponse qui aura pu être étalonné sur les périodes antérieures. Combiné avec l'information externe, on obtiendra des poids transversaux $w_k^{t^*}$ permettant l'analyse des données transversales, les estimations ponctuelles et les estimations de variance. La non-réponse partielle pourra être imputée grâce à toutes les variables auxiliaires disponibles, que celles-ci soient transversales où longitudinales. C'est le cas, par exemple, quand on utilise une imputation par ratio du style :

$$\hat{y}_k^{t^*} = y_k^{t^{**}} \left(\frac{\bar{y}^{t^*}}{\bar{y}^{t^{**}}} \right)_r$$

où le ratio est calculé sur les répondants communs aux époques t^* et t^{**}.

Ainsi, l'analyse transversale est de nature parfaitement classique. Elle se caractérise seulement par l'abondance (si le panel est ancien) et la nature de l'information auxiliaire disponible.

L'analyse longitudinale ne nécessite pas de nouvelles techniques. La non-réponse totale sera évidemment repondérée en fonction de l'information liée à la base où de l'information externe actuelle (attention, cependant, celle-ci doit être vue comme relative à la population U, et pas à la population U_t. Autrement dit tout calage "longitudinal" sur une donnée externe X_t relative à U_t s'écrira $\sum_s E_k^{(t)} x_k^t w_k = X_t$).

Le problème spécifique réside dans le fait que la non-réponse totale à une vague du panel doit être considérée comme une non-réponse partielle. Si on s'en tient aux habitudes, ce type de non-réponse est compensée par imputation, en utilisant toute l'information antérieure ou postérieure (interpolation à la date considérée).

Une constatation générale vient tempérer quelque peu cette façon de voir. Une grande partie de la non-réponse totale après la première vague est définitive (en tous cas on n'a pas observé de retour dans le champ des répondants à l'époque t*!). C'est le phénomène d'érosion (in English attrition) d'un panel. Cette érosion est particulièrement forte après la première (et parfois aussi la seconde) vague.

Il paraît peu esthétique de procéder à des imputations massives d'unités présentes à seulement une ou deux époques de l'étude longitudinale. On peut procéder de la façon suivante. Supposons, pour fixer les idées, qu'on ne décide d'imputer la non-réponse totale qu'à partir de la troisième vague, la seconde étant en quelque sorte celle où on considère que la fidélisation au panel se stabilise. Les répondants au panel longitudinal seront alors, par convention, ceux de la seconde vague (échantillon r).

Il s'agit de ne pas jeter totalement l'information apportée par les répondants $\left(\text{échantillon } r_1 \text{ , } avec\ s \supset r_1 \supset r\right)$ de la première vague. Pour ce faire, on pourra fabriquer un jeu de pondérations longitudinales (relatives à l'échantillon r) compatible avec l'information apportée par r_1. On y parviendra en ajoutant aux équations de calage estimant les paramètres du modèle de non-réponse - qui sont de la forme $\sum_r w_k x_k = X$ - de nouvelles équations :

$$\sum_r w_k \mathbf{z}_k = \sum_{r_1} w_k^{(1)} \mathbf{z}_k = \mathbf{Z}.$$

Là-dedans, $\mathbf{Z}$ désigne le vecteur des variables de la première vague dont on désire respecter l'information et $w_k^{(1)}$ le jeu des pondérations transversales adoptées pour l'exploitation de la première période. $\mathbf{z}_k$ peut contenir, par exemple, des indications relatives aux catégories d'activité à l'époque 1 si l'érosion se différencie surtout selon

ces variables. Cette technique se généralise sans problème technique particulier à la prise en compte de plusieurs périodes. Il faudra simplement prêter attention aux choix des statistiques sur lesquelles on décide de caler.

Conclusion provisoire :

L'exploitation des données collectées dans un panel se ramène à des techniques bien connues que ce soit en matière de correction de la non-réponse ou de choix d'estimateur. Le problème réside surtout dans les choix des contraintes conduisant à une certaine forme de calage ; ce choix peut s'avérer un peu délicat si on introduit des contraintes générales de cohérence entre les exploitations longitudinales et transversales.

6 - Quelques problèmes liés à l'échantillonnage

Nous avons ramené, de façon un peu formelle il est vrai, le problème de l'échantillonnage pour panel à celui de l'élaboration d'un plan de sondage p(s) sur la population d'étude $U = \underset{t \in T}{U} U_t$. Si celle-ci est connue à l'instant initial d'échantillonnage, il n'y a pas de problème (Exemple : Population des personnes de 18 ans et plus ayant vécu en France jusqu'à cet âge ; panel avec un horizon de 8 ans ; la base de sondage est un recensement réalisé l'année qui précède la mise en route du panel).

Généralement, malheureusement, ce n'est pas le cas. On utilise une base de sondage pour la population U_{t_0} avec un plan $\left(p_0(s) ; s \subset U_{t_0}\right)$ qui permet de tirer l'échantillon longitudinal, celui qu'on utiliserait seul si on s'intéressait à une analyse par cohorte. Cet échantillon verra sa taille diminuer au cours des périodes successives par le jeu de la "mortalité" (= sortie de champ en général) et de l'érosion (qu'il faudrait arriver à distinguer le mieux possible même si ce n'est pas simple : un "mort" est aussi un non-répondant définitif !). Maintenant, les périodes successives $(t, t + \Delta t)$ seront, pour l'échantillonnage, traitées comme des strates avec une base de sondage spécifique $\beta_t^{\Delta t}$ (quand cela est possible !) . La limite de ce système est celui de l'enregistrement continu avec enrichissement continu de la base. Chaque unité sondable arrive dans la base à un instant spécifique. Elle peut être mise en réserve jusqu'au tirage (cas stratifié) ou être échantillonnée tout de suite. La seule technique utilisable alors est celle de l'échantillonnage Poissonnien où l'unité est incluse dans l'échantillon avec une probabilité π_k qui ne dépend que de ses caractéristiques propres et de l'information auxiliaire. Les pratiques qui consistent à rééchantillonner en fonction de l'érosion du panel (pour conserver un nombre de répondants fixe d'une vague à l'autre) n'ont aucune justification liée à la représentativité de

l'échantillonnage ; elles conduisent simplement à attribuer à la strate $\beta_t^{\Delta t}$ un poids d'extrapolation lié au taux d'érosion des vagues précédentes.

Remarque :

Le fait de compléter le panel par un nouvel échantillon probabiliste dans $\bigcup_{t+\Delta t}$ est de nature différente et complique singulièrement les choses. Quand cet échantillon est exploité avec l'échantillon panel, on utilise généralement des pondérations qui supposent les deux échantillonnages indépendants (meilleur estimateur linéaire sans biais). En effet, généralement, les probabilités d'inclusions ne sont pas toujours calculables facilement. De fait si $\tilde{s} = s \cup s'$ on aura $\tilde{\pi}_k = P_r(k \in \tilde{s}) = P_r(k \in s) + P_r\left(k \in s' \middle| k \notin s\right)$. Généralement, si k vient effectivement de s, le second terme sera malaisé à récupérer, et inversement si k vient de s'.

7 - L'échantillonnage indirect

De fait, on échantillonne rarement les personnes (pour un panel ou pas !) à partir d'une base de sondage de personnes. Dans la pratique française du Panel Européen, l'échantillonnage est réalisé à partir de logements, qui permettent d'attraper les personnes, et, de façon intermédiaire, les ménages, qui sont un ensemble de personnes habitant un logement siège d'une résidence principale (sur l'articulation entre ces termes et une approche rigoureuse de ces définitions on pourra se reporter à [J-C.Deville (1988)], dans un ensemble de documents qui ne fut pas jugé digne d'entrer dans le sanctuaire de *Données Sociales*, où les données sont vraiment considérées comme données).

Ces logements sont échantillonnés dans l'échantillon-maître (EM) - ce qui permet une extrapolation à l'ensemble des logements recensés en mars 1990 - et dans la Base de Sondages des Logements Neufs (BSLN), qui est un panel des autorisations de construire décernées depuis 1987. Ce panel est lui-même échantillonné sur une base géographique qui a le bon goût, malgré diverses chicaïas administratives et la dérive des continents, d'être fixe dans le temps.

Autrement dit, le panel de logement qu'est l'échantillon-maître complété de la BSLN, permet d'entretenir un panel de personnes. Voyons comment.

La vague initiale sera échantillonnée à partir d'un échantillon habituel de logements. Chaque vague successive - théoriquement - doit être enrichie d'un échantillon de logements construits depuis la vague précédente. On construit ainsi un panel de logements (sous-panel de la base EM + BSLN !). Ce panel de logements a la

propriété d'être extrapolable transversalement à toute époque d'exploitation potentielle du panel de personnes. On obtient donc un échantillon extrapolable de personnes de la façon suivante :

– A t_0 (population longitudinale), l'échantillon est constitué de toutes les personnes du champ (condition d'âge éventuelle) trouvée dans les logements de l'échantillon de logements L_{t_0}. Ces personnes seront suivies jusqu'à l'horizon du panel, quel que soit le logement qu'elles occupent aux dates successives d'enquêtes.

– Pour toute date ultérieure t , soit L_t l'échantillon du panel de logements à l'époque t (échantillon initial L_{t_0} + logements neufs) ; on inclut dans le panel de personnes celles qui sont entrées dans le champ ("naissances", en pratique immigrés) depuis la date d'enquête antérieure.

Comme la majorité des personnes ne déménagent pas entre deux périodes d'enquêtes, cette méthode est assez économique.

Remarque :

En fait les "naissances" de bébés sont traitées un peu différemment. On utilise la particularité qu'ils ont d'avoir une mère, qui, éventuellement, peut avoir la chance de faire partie du panel.

8 - Echantillonnage indirect et "panels" de ménage

Ce qu'on vient d'analyser est un cas particulier d'échantillonnage indirect dont on va donner maintenant l'ébauche d'une théorie un peu plus générale et dont les applications sont multiples et même innombrables.

Une population U est échantillonnée selon un plan de sondage $(p(s)\,;\, s \subset U)$ autorisant à faire des statistiques grâce à des probabilités d'inclusion π_k et des techniques d'estimation comme cela a déjà été évoqué. On cherche à atteindre une population V d'individu courant i. Une matrice $A = \{a_{ki}\}$ de nombres positifs ou nuls relie ces deux populations. Pour que l'affaire marche bien, comme on le verra, il faut que la matrice A soit assez creuse, les a_{ki} non nuls étant rares.

Echantillonnant l'unité k de U, on enquêtera toutes les unités i de V telles que $a_{ki} > 0$. Les a_{ki} sont collectables auprès de l'unité k (et ne sont pas nécessairement connus dans toute la base de sondage). Lors de l'enquête auprès de l'unité i de V on collecte

également les a_{ki} positifs pour i fixé. Formellement, donc s_U est tiré dans U selon le plan p.

On en déduit un échantillon $s_V = \{i \in V\ ;\ \exists k \in s_U \text{ et } a_{ki} > 0\}$.

On collecte :
- les $a_{ki} > 0$ pour $k \in s_U$
- les $a_{ki} > 0$ pour $i \in s_V$

On suppose que pour tout i de V il existe au moins une unité k de U telle que $a_{ki} > 0$ de façon à ce que toute la population V puisse être ainsi attrapée.

Donnons quelques exemples bien connus :

Exemple 1 :
U est la population des logements k, $V = \{i\}$ est un ensemble de personnes ; $a_{ki} = 1$ si M^r ou M^{me} i habite le logement k. Sinon $a_{ki} = 0$.

Exemple 2 :
$V = \{i\}$ peut aussi être un ensemble de ménages et alors $a_{ki} = 1$ si k est la résidence principale du ménage i (et vaut zéro sinon). Dans ce cas V est une partie de U puisqu'on identifie ménage et résidence principale.

Exemple 3 :
Sondage en grappes (c'est la généralisation de l'exemple 1).

Exemple 4 :
On admet que tout enfant de moins de 10 ans vit avec une personne majeure de plus de 18 ans inscrite sur la "Liste des Personnes Majeures". On tire un échantillon dans cette liste pour attraper des petits enfants. La relation est $a_{ki} = 1$ si l'enfant i habite avec la grande personne k. On notera que à chaque k, peuvent être associés plusieurs i et inversement.

Exemple 5 :
Un ensemble d'entreprises $V = \{i\}$ est possédé par des actionnaires k qu'on attrape parce qu'ils paient des impôts ; a_{ki} est le montant du capital de i possédé par k. On a ici un exemple de nature numérique particulièrement intéressant car on peut faire des transformations comme par exemple ne pas prendre en

compte les a_{ki} inférieurs à un certain seuil : $a'_{ki} = a_{ki}$ *si* $> a$ et égal à 0 sinon ; ou aussi $a''_{ki} = 1$ si $a_{ki} > a$, etc.

Exemple 6 :
Entretien d'un panel de logements.
La démographie des logements est assez simple. Outre la phase prénatale (autorisation, mise en chantier, ...), nous sommes préoccupés par l'achèvement, la destruction et deux transformations assez simples : la fusion et l'éclatement (la recomposition peut également s'envisager sans trop poser de problème). Dans tous les cas, si nous nous intéressons à la population "avant" et "après" l'événement, la matrice A est définie par $a_{ki} = 1$ si le logement k "avant" participe au logement i après. Pour ce qui concerne la gestion des identifiants une règle arbitraire d'héritage en cas de fusion, ou de "déclinaison" en cas d'éclatement, peut être appliquée selon ce qu'on entend par logement identique "après" et "avant".

Le procédé peut s'itérer pour attraper une troisième population W, d'individu courant j, liée à V par une matrice $B = (b_{ij})$, d'éléments positifs ou (le plus souvent) nuls. Sur ce principe, on peut aussi faire des enquêtes en deux phases en échantillonnant dans l'échantillon s_v. Bref la procédure est assez riche de possibilités. Avant de théoriser là-dessus, annonçons le dernier exemple.

9 - Comment faire des statistiques sur les ménages dans un panel, suivi de, comment enrichir un panel sans trop se fatiguer

La principale application de l'échantillonnage indirect est la possibilité d'attraper des ménages grâce aux individus. Dans la pratique française c'est automatiquement réalisé lors de la première vague puisque nous démarrons sur un échantillon en grappes. Dès la vague suivante, on doit se poser le problème à cause de la démographie bizarroïde et mal connue des ménages. De plus, les individus de ces ménages "transversaux" (comme on dit !) sont eux-mêmes objet d'enquête car l'information qu'ils peuvent apporter s'avère non seulement non inutile mais surtout pas chère à collecter.

Les matrices qui relient les "personnes-panels" aux ménages et aux individus de ces ménages sont les suivantes :

- $a_{ki} = 1$ si la personne-panel k est dans le ménage i
- $b_{ij} = 1$ si la personne j (de U_t) appartient au ménage i

– $c_{kj} = \sum_i a_{ki}\, b_{ij}$ $(C = AB)$ si la personne j appartient au ménage de la personne-panel k.

On pourra remarquer que dans B on peut introduire des caractéristiques particulières des personnes permettant une exploitation (cf. paragraphe 10) sur des populations particulières. On peut aussi enrichir le panel sans trop se fatiguer en chaînant le procédé et en panélisant les individus ainsi attrapés. Ceci demande un certain doigté pour une exploitation longitudinale. Pour faire du transversal, en revanche, comme on le verra, tout va bien (de même, ce qui est assez chouette, que pour faire des statistiques d'accroissement net !).

La façon dont on traite les bébés est une autre application, fort utile. L'accroissement de population $U_{t+\Delta t} - U_t$ comporte deux parties : les immigrants (hors champ à t mais vivants) et les bébés (nés entre t et t + Δt). On a vu comment on pouvait attraper les immigrants par la technique du panel de logements. On pourrait faire la même chose avec les bébés mais il serait un peu gênant (quoique ?) de sélectionner seulement le bébé d'une famille qui vient de s'installer dans un logement panel, ou de suivre les personnes du ménage rien qu'à cause de ce foutu bébé. L'alternative consiste à capturer les bébés par leurs parents. Donc, dans la matrice A, on aura $a_{ki} = 1$ si l'individu "panel" (ou pas !) k est parent du bébé i. Pour des raisons d'incertitude génétique et de tradition des démographes, on aura en fait $a_{ki} = 1$ si l'individu "panel" (ou pas) k est la mère du bébé i. Le bébé en question peut alors être panélisé avec un poids qui va apparaître dès le paragraphe suivant.

10 - Pondération et échantillonnage indirect. La méthode de partage des poids

– Pondération : On s'intéresse au total $\sum_V y_i = Y$ d'une variable y de la population V. Si on note 1_v le vecteur avec des 1 pour chaque indice i et $y = (y_i)$ on peut écrire $Y = I_V' \,.\, y$ (produit scalaire). Posons $a_{.i} = \sum_{k \in U} a_{ki}$. On a l'identité $Y = \sum_{k,i} \frac{a_{ki} y_i}{a_{.i}}$. On suppose - c'est essentiel n'est-ce pas - que $a_{.i}$ a été collecté à l'enquête. Dans le cas des personnes des ménages transversaux, $a_{.i}$ est tout simplement le nombre de personnes panélisables présentes dans le ménage dont i fait partie (et donc le nombre de personnes du ménage tout court si le panel est sans restriction, d'âge par exemple).

La variable $z_k = \sum_{i \in V} a_{ki} \frac{y_i}{a_{.i}}$ est donc définie pour tout k de U et mesurée pour tout k de s_U. On a, bien évidemment, $\sum_U z_k = \sum_V y_i$. Notons qu'on peut écrire $z = A\, diag\left(A' I_U\right)^{-1} y = \tilde{A} y$ si on aime les notations matricielles. Si on dispose de pondérations w_k sur les individus panels (par exemple les poids de l'estimateur de Horvitz-Thompson), on dispose aussi d'un estimateur de

$$Z = \sum_k z_k = I_U' z = \left(I_U'.A\right) diag\left(A' I_U\right)^{-1} y = I_V'.y = Y \text{ par } \hat{Z} = \hat{Y} = \sum_{s_U} w_k z_k = w'.z$$

si on note w le vecteur des w_k.

Sous cette forme il est donc clair qu'on obtient un estimateur de $Z = Y$ sans biais si les poids sont sans biais et dont on sait estimer la variance si on sait le faire pour les poids w_k.

On peut, de même, estimer toute fonction de totaux (estimateur par substitution) ainsi que la variance de cette statistique (linéarisation). Bref tous nos problèmes sont résolus.

On peut aller plus loin, et profiter d'une information auxiliaire que ce soit au niveau de la population U (ce qui va de soi) mais aussi au niveau de la population V, et bien entendu, des deux simultanément. Voyons d'abord la forme opérationnelle (fichier de dépouillement) de cette méthode. Transformons l'estimateur $\hat{Z}$.

$$\hat{Z} = \sum_{s_U} w_k z_k \sum_{s_U} w_k \sum_{i \in V} a_{ki} \frac{y_i}{a_{.i}} = \sum_{s_V} y_i \sum_{s_U} \frac{w_k a_{ki}}{a_{.i}} = \sum_{s_V} w_i^* y_i$$

Matriciellement :

$$\hat{Z} = w' \tilde{A} y = w^{*'} y \text{ avec } w^* = w' \tilde{A}.$$

On a $w_i^* = \frac{1}{a_{.i}} \sum_k w_k a_{ki}$ où, évidemment, la somme ne porte que sur les $w_k a_{ki}$ non nuls, c'est-à-dire sur les individus k "rattachés" à i par la positivité des a_{ki}. Dans le cas des "individus transversaux" du panel, la somme est celle de tous les poids des "personnes panels" appartenant au ménage de i, $a_{.i}$ est le nombre de personnes (du champ) de ce ménage. D'où le terme de partage des poids [Ernst (1989)] donné à cette méthode longtemps considérée comme empirique. On remarquera que si le ménage de V a la même composition que le ménage de U qui "pointe" sur lui, les individus ont pour poids la moyenne des w_k (ce qui ne change rien si ces poids

sont égaux). En revanche les personnes attrapées par une seule personne-panel se voient contraintes à partager le poids de cette dernière.

Indiquons enfin comment on peut profiter de l'information auxiliaire présente au niveau de la population V. Nous admettons que celle-ci se présente sous la forme d'un vecteur X de totaux connus sur V. Le respect de l'information nous demande donc de satisfaire l'équation de calage :

$\sum_{s_V} w_i x_i = X = w' \tilde{A} \underline{x} = \sum_{s_U} w_k \left(\sum_i \tilde{a}_{ki} x_i \right)$ *où* $\underline{x}$ est la matrice qui empile des x_i.

On est donc ramené à un problème standard de calage, si on le désire. Le partage des poids sera alors un partage des poids calés sur s_U. On pourrait concevoir un calage direct des poids w_i^* eux-mêmes. La question de la cohérence avec les calages sur U doit alors s'étudier, mais, en ce qui me concerne, plutôt un autre jour.

14 - Conclusion

Que ce soit dans les aspects liés à l'échantillonnage ou dans les aspects liés à l'estimation, les enquêtes par panels ne posent essentiellement pas de problèmes nouveaux. La difficulté réside surtout dans la mise en œuvre des procédures "classiques", dans le fait d'identifier la nature exacte des problèmes. En particulier, certaines difficultés surviennent quand on cherche à mettre en cohérence des exploitations relatives à des ensembles de dates différents (transversal et longitudinal par exemple).

Enfin, la technique d'échantillonnage indirect, qu'on pratique déjà sans le savoir dans les enquêtes ponctuelles, devient un outil essentiel dans la statistique de panels. Il faut savoir identifier les situations où elle s'avère utile et la mettre en œuvre à bon escient. Elle permet toutes les exploitations transversales imaginables et peut même permettre, si on n'est pas trop difficile, d'entretenir un panel sur une durée indéfinie.

Références et bibliographie :

BINDER D.A, "Longitudinal surveys : why are these surveys different from all other surveys ?" IASS/IAOS Satellite Meeting on Longitudonal Surveys, Jerusalem, August 27-31, (1997)

CHAMBAZ C, SAUNIER J.M, VALDELIEVRE H, "Méthodologie du panel européen de ménages : exploitation des données de la vague 2 du fichier français", Insee, Direction des Statistiques Démographiques et Sociales, document de travail F 9715 (1997)

COTTON F, HESSE C, "Méthodes d'échantillonnage pour l'enquête annuelle d'entreprises, Actes des JMS de 1991, Insee Méthodes, vol. 29-30-31, (1992)

DEVILLE J.C, *Structure des familles : résultats de l'enquête de 1962,* Collections de l'Insee, Vol. D, 13-14, (1972)

DEVILLE J.C, "Peut-on croire aux enquêtes ?" dans *Construire les données sociales,* Collections de l'Insee, Vol. M 128, pp. 15-22, (1988)

ERNST L.R, "Weighting issues for longitudinal and family estimates", dans *Panels Surveys,* edited by KASPRZYK D, DUNCAN G, KALTON K, and SINGH M.P Wiley (1989)

HESSE C, "Tirage, rotation, retirage d'un panel stratifié de taille fixe : la méthode panastra", Insee, Direction des Statistiques d'Entreprises, (1994)

HOLT D, SKINNER C.J "Components of change in repeated surveys", International Statistical Review, Vol. 57, pp. 1-18, (1989)

KALTON G, CITRO C.F, "Enquêtes par panel : ajout d'une quatrième dimension" Techniques d'Enquête, Vol. 19, pp. 217-227, (1993)

LAVALLEE P, Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids" Techniques d'Enquête, Vol. 21, pp. 27-35, (1995)

LAVALLEE P, "Représentativité et pondération dans les enquêtes longitudinales", Université de Caen (1996)

LEROY-LADURIE E, *Histoire du climat depuis l'an 1000,* Flammarion (1997)

BINDER D.A, "Longitudinal surveys : why are these surveys different from all