N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2004 : Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

2004

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2004 de Statistique Canada
Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

EXAMEN DES STRATÉGIES EMPLOYÉES DANS LES ENQUÊTES-ÉTABLISSEMENTS DE L'ONS DANS LE CAS DES POPULATIONS RARES ET DIFFICILES À JOINDRE

Paul Smith et John Perry[1]

RÉSUMÉ

Les enquêtes-établissements de l'Office for National Statistics (ONS) portent sur un grand nombre de domaines et plusieurs visent des populations pour lesquelles on ne dispose pas de bonnes bases de sondage ou s'emploient à mesurer des caractéristiques rares dans la population des entreprises. L'ONS aborde ces difficultés de diverses manières. D'abord, il dresse des registres satellites individuels pour des enquêtes (dans le secteur financier, par exemple) en comptant sur une source administrative. Ces sources peuvent ne pas avoir une couverture complète, mais elles seront généralement d'une bonne qualité si elles viennent d'organismes de réglementation. Une autre possibilité est de construire un panel contenant l'information historique. Cela comporte des défis d'échantillonnage puisqu'il est nécessaire d'avoir une couverture supplémentaire pour la population qui est en dehors du panel.. Il reste que, avec cette orientation, l'actualisation du panel et le traitement des « valeurs aberrantes » peuvent influer sur les estimations. Une autre possibilité est d'utiliser une question filtre d'une enquête plus générale par laquelle on constituera un échantillon au second degré. Le principe d'un échantillonnage à deux degrés relève d'une théorie mieux élaborée, mais des questions d'ordre pratique subsistent quant aux décalages de période entre les deux degrés et au mode de traitement des observations atypiques. Cela peut aussi être une façon d'étendre ou d'actualiser le panel. Dans ce document, nous passerons en revue les méthodes en question et les défis que présente leur application et citerons des exemples tirés des enquêtes-établissements de l'ONS.

MOTS CLÉS : Caisses de retraite, commerce international des services, échantillonnage à deux degrés, registres satellites.

1. INTRODUCTION

1.1 Enquêtes portant sur des caractéristiques rares

Dans les administrations publiques, les besoins d'information sur l'activité des établissements (« établissements » et « entreprises » seront synonymes dans notre exposé) sont nombreux et visent toutes sortes de branches d'activité économique. Dans certains de ces secteurs, la population est bien connue, les entreprises étant tenues de s'inscrire à un régime administratif. Au Royaume-Uni, les principales sources administratives sont les régimes d'imposition de la valeur ajoutée (TVA) et de l'impôt à la source sur le revenu (PAYE), ainsi que la Companies House. Nous décrivons sommairement à la section 1.2 la façon de constituer la population type des enquêtes-établissements de l'ONS.

Un certain nombre d'activités ne relèvent cependant pas d'un ou de plusieurs régimes administratifs, auquel cas les populations cibles seront moins bien connues et donc plus difficiles à enquêter. Pour ces cas, nous utilisons une combinaison de sources pour constituer la population à échantillonner et nous devons employer un certain nombre de stratégies garantes de la meilleure couverture possible de cette population.

Une autre difficulté se présente lorsque la population d'établissements est bien définie, mais que la caractéristique à mesurer dans l'enquête est rare dans la population. Les méthodes habituelles d'enquête produiront inévitablement en pareil cas des estimations avec une grande variance que l'on pourra difficilement utiliser comme indicateurs

[1] Paul Smith, Office for National Statistics, chemin Cardiff, Newport, NP10 8XG, Royaume-Uni (paul.smith@ons.gov.uk); John Perry, Office for National Statistics, chemin Cardiff, Newport, NP10 8XG, Royaume-Uni (john.perry@ons.gov.uk).

économiques. Pour réduire la variabilité, il faudra se reporter à d'autres renseignements sur la présence de la caractéristique d'intérêt dans la population.

1.2 Identification de la population d'entreprises au Royaume-Uni

L'identification de la population d'entreprises du Royaume-Uni se fait à des fins statistiques à l'aide d'un registre interministériel d'entreprises appelé IDBR (Inter-Departmental Business Register; Perry, 1995). Celui-ci vient de trois sources administratives :

- Régime de la taxe à la valeur ajoutée (TVA) auquel doivent obligatoirement s'inscrire les entreprises dont le chiffre d'affaires est supérieur à une valeur seuil fixe (58 000 £ à compter d'avril 2005). Les données d'inscription indiquent à la fois la catégorie de l'entreprise par un code descriptif d'une ligne de son activité et une estimation – fournie par l'entreprise même – de son chiffre d'affaires annuel. Pour certaines activités (des caisses de retraite, par exemple), l'inscription n'est pas obligatoire, mais volontaire.
- Régime d'impôt à la source sur le revenu (PAYE). C'est un régime de perception de l'impôt sur le revenu et des cotisations d'assurance nationale (ce qui comprend les contributions des employeurs). Tout employeur dont les salariés gagnent plus qu'une certaine valeur seuil (qui est aujourd'hui d'environ 5 000 £) est tenu de s'inscrire au régime PAYE.
- Companies House. Les entreprises qui désirent se constituer en société ont l'obligation de s'inscrire à la Companies House. Elles figurent dans un registre public dont l'information sert aussi à l'élaboration de registres. Les entreprises constituées en entreprise (généralement des sociétés ouvertes à responsabilité limitée) remettent d'année en année leurs états financiers à la Companies House, mais l'information ainsi réunie est moins à jour que celle des deux autres sources.

Dans bien des cas, on pourra reconnaître une entreprise par les trois sources, mais dans certains secteurs d'exonération du régime TVA, il n'y aura comme source que le régime PAYE, ce qui posera tout particulièrement un problème dans le cas du secteur financier. Comme certaines entreprises financières se caractérisent par un nombre de salariés relativement bas et peut-être par une structure complexe, il peut s'avérer difficile de repérer uniquement par les chiffres d'emploi les unités d'enquête à viser par une telle source administrative.

1.3 Aperçu

À la section 2, il sera question des registres satellites et des problèmes que posent leur mise à jour et leur alignement sur le registre principal. À la section 3, nous décrirons brièvement les méthodes d'échantillonnage à deux degrés pour les enquêtes auprès de populations rares. À la section 4, nous nous attacherons à la question de l'obtention de registres en utilisant l'approche du panel. À la section 5, nous confronterons les différentes méthodes et ferons voir les secteurs où un complément de recherche s'impose.

2. REGISTRES SATELLITES

2.1 Intégration d'enquêtes à partir de registres

Ces dix dernières années, l'ONS s'est servi de l'IDBR comme base de sondage pour ses principales enquêtes auprès des entreprises. L'avantage en est qu'une base de sondage unifiée permet la coordination des enquêtes; on se trouve alors à coordonner les échantillons et on s'assure que les diverses enquêtes couvrent l'éventail des activités dont fait état le registre. On évite ainsi les problèmes antérieurs reliés à l'utilisation de bases de sondage distinctes pour différents secteurs, des entreprises pouvant être comptées en double ou manquées dans un éventail d'enquêtes.

Ces difficultés subsistent néanmoins dans les secteurs qui sont mal couverts par les données administratives. Pour prendre un exemple précis, la mesure de l'exploitation et des flux financiers des caisses de retraite est un sujet de grande actualité au Royaume-Uni, mais bien des caisses sont gérées de telle manière qu'elles peuvent ou non être individuellement caractérisables comme entreprises. Ajoutons que le régime TVA applique des règles particulières de traitement de la capitalisation des régimes de retraite, ce qui donne lieu à un certain nombre d'exonérations de

TVA (HMCE, 2002). Il n'est donc pas possible de reconnaître toutes les activités des caisses de retraite par les sources qui forment l'IDBR. On dispose de deux autres grandes sources d'information qui sont propres l'une et l'autre à cette branche d'activité. L'Occupational Pensions Regulatory Authority prescrit l'inscription OPRA de tout régime de retraite. Les régimes en question ne sont toutefois pas groupés en caisses et leur dénombrement laisse quelque peu à désirer. Ainsi, les administrations locales ne sont pas tenues de s'inscrire au régime. L'autre source est la National Association of Pension Funds (NAPF), association sectorielle qui représente largement – mais non entièrement – cette branche d'activité au Royaume-Uni et qui est renseignée sur les caisses membres, mais sans savoir précisément quels régimes composent ces caisses.

Soucieux d'avoir une couverture maximale à un coût minimal pour lui et pour les entreprises qui répondent aux questionnaires, l'ONS se sert actuellement du registre NAPF comme source. Son échantillonnage s'est fait au niveau des caisses de retraite, bien qu'il soit en train de revoir sa stratégie. Pour que le tout fonctionne, il lui a fallu élaborer un registre satellite.

2.2 Registres satellites

Un *registre satellite* est un registre fondé sur une source indépendante, mais qui se trouve lié à un registre principal ou « gravite » pour l'essentiel autour de ce registre. Le principe est simple, parce qu'on fait le plus de recoupements possible entre le registre satellite et le registre principal dans ce qui se veut la plus grande synchronicité de mise à jour des deux registres.

Dans le cas des régimes de retraite, la source NAPF sert à dresser un registre des caisses et de leurs adresses. On dénombre en gros 900 de ces caisses. Dans la mesure du possible, on procède à un appariement par adresse et mode d'activité avec les entreprises figurant déjà à l'IDBR. Ce processus laisse un certain nombre de caisses absentes de l'IDBR, c'est-à-dire non inscrites au régime TVA ni au régime PAYE. Ces caisses sont alors incluses dans la population cible de l'enquête.

Comme l'IDBR sert à la fois de moyen de contrôle de l'échantillon et de base de sondage, on crée des unités pour les entreprises non appariées. Unités appariées et unités non appariées sont marquées comme unités à viser directement par l'enquête sur les caisses de retraite. Elles forment le registre satellite comme sous-ensemble de la base d'information normale du grand registre. La sélection et la tenue à jour de l'échantillon peuvent alors se faire comme d'habitude avec la population réduite définie par ces « entreprises du secteur des retraites ».

La mise à jour prend alors deux formes : mise à jour habituelle des sources administratives pour le registre entier avec les changements possibles de structure d'entreprise mère; mise à jour en fonction des changements de structure de caisse de retraite qui sont périodiquement tirés du régime NAPF. Dans le premier cas, on se trouve à confirmer les activités (c'est-à-dire à vérifier la structure d'entreprise) et, dans le second, à mettre à jour les unités par rapport au registre en les appariant ou en créant ou modifiant des unités. Les sources peuvent toutefois ne pas présenter les mêmes caractéristiques. Ainsi, la source NAPF comporte environ 5 % de créations en deux ans et la source IDBR, 7,5 % chaque année pour les entreprises des secteurs des retraites et des assurances.

La source NAPF renseigne aussi sur l'actif total des caisses de retraite, caractéristique qui devient une variable auxiliaire tant dans la stratification que dans l'estimation. On l'obtient directement des mises à jour de cette source.

2.3 Questions de dénombrement et d'estimation dans le cas des registres satellites

Comme nous l'avons indiqué, un registre satellite peut être là pour compléter les unités du registre principal des entreprises et fournir des variables supplémentaires pour la stratification ou l'estimation. Il se peut néanmoins que les registres satellites ne soient pas complets.

Pour reprendre l'exemple des régimes de retraite, notons que le registre NAPF est en sous-dénombrement et qu'une comparaison antérieure avec la source OPRA semble indiquer qu'il manque 1,6 million de cotisants aux régimes de retraite sur un total connu de 12,2 millions (proportion de 13 %). Une telle comparaison est compliquée parce que les unités prises en compte dans les deux sources ne sont pas en concordance. Pour compenser quelque peu le sous-

dénombrement caractéristique de la source NAPF et tenir compte de la partie non observée de la population, nous prévoyons ajouter un facteur d'inflation aux estimations relatives aux caisses de retraite. Ce facteur de pondération est actualisé tous les deux ans.

La sous-couverture relativement élevé de NAPF nous a amenés à nous interroger dans une certaine mesure sur l'utilisation de la source OPRA comme meilleure base de sondage. Une stratégie possible serait de se reporter à cette source comme base principale pour la collecte de données, mais de faire de la source NAPF une base secondaire comme moyen de constatation des lacunes de la base OPRA en vue d'améliorer l'estimation de l'activité globale des régimes de retraite.

3. ÉCHANTILLONNAGE ET ESTIMATION À DEUX DEGRÉS

3.1 Questions filtres

Dans ce contexte, une question filtre est une question que l'on pose normalement dans une enquête comportant un grand échantillon et une bonne couverture (activité d'enquête au premier degré) et par laquelle on constate qu'une entreprise présente une certaine caractéristique et fait, par exemple, de la recherche-développement (R-D). Avec cette question, on relève la présence d'une telle activité pour une grande diversité d'entreprises à un coût relativement bas et, en particulier, à un coût d'observation peu élevé pour l'entreprise qui doit répondre à l'enquête. Au second degré et en suivi, on pose des questions plus détaillées à un sous-échantillon dégagé par la question filtre au premier degré. C'est ce qu'on pourrait appeler une « question filtre cachée », car l'entreprise qui répond à la première question ne sait pas que, à la suite, on pourrait solliciter d'autres renseignements, bien que l'ONS ait soin de donner l'assurance et la garantie publique que tous les renseignements ainsi recueillis le seront uniquement à des fins statistiques.

3.2 Approche à deux degrés

Comme cette orientation s'insère tout à fait dans une structure d'enquête à deux degrés, les méthodes de conception et d'analyse de telles enquêtes sont bien connues. Il y a deux types de poids d'échantillonnage, l'un projette les réponses de la deuxième phase au total de l'échantillon de la première phase, et l'autre projette la première phase à la population. Il existe en fait diverses façons d'utiliser des totaux auxiliaires et des estimations au premier degré, ainsi que le décrivent Estevao et Särndal (2002). En cas de question filtre par oui ou non, le cas (B2) de ces auteurs devient l'expression type $\sum w_I \sum w_{II} y_{ki}$, mais en cas de question filtre plus détaillée comme sur l'activité totale dans la catégorie visée (volet « commerce international de services » de l'enquête ABI (Annual Business Inquiry) de l'ONS depuis 2000, par exemple), w_{II} peut servir à calibrer y_{kII} par rapport à $\sum w_I y_{kI}$ (cas (A3)). Dans la pratique, l'ONS n'a pas adopté ce mode de calcul pour le volet « commerce international de services ». Pour le rapprochement des données, on se sert plutôt du total estimé au premier degré $\sum w_I y_{kI}$ afin de déceler les anomalies dans les sources de données.

Il est révélateur de considérer les coûts et les avantages relatifs du traitement à deux degrés par rapport à la solution habituelle consistant à poser des questions détaillées au départ. L'avantage premier est que le coût est relativement bas, les questions détaillées n'étant posées qu'aux entreprises dont on sait qu'elles exercent une activité qu'on entend mesurer. Les coûts baissent pour l'ONS et les coûts d'observation diminuent un peu pour les enquêtés, puisque les entreprises qui n'exercent pas l'activité en question n'ont pas à répondre aux questions détaillées. Trois coûts principaux sont à mettre en balance avec cet avantage. Premièrement, il se peut qu'il y ait relativement peu d'unités de l'échantillon du premier degré qui présentent la caractéristique visée et, en cas de non-réponse, toutes ne seront pas reconnues. Il y a, par exemple, une entreprise sur dix parmi les plus petites qui fait des virements internationaux (volet « commerce international de services ») comparativement au tiers des entreprises les plus grandes. Autre aspect : on aura beau bien reconnaître cette caractéristique, mais les entreprises pourraient ne pas effectuer de virements internationaux tous les ans. Ainsi, une équipe de football pourrait en faire une seule fois en quelques années. En fait, 6 des 10 réponses d'entreprises reconnues à ce titre au premier degré sont devenues des

réponses néant au second degré en 2001. Si l'échantillon ne suffit pas à une mesure, nous pouvons le compléter par un échantillonnage d'entreprises non visées au premier degré, mais nous ne saurons pas lesquelles ont l'activité en question, d'où le risque de gaspiller de l'argent et de produire des estimations ayant une variance importante.

Deuxièmement, il peut y avoir un décalage temporel entre les deux degrés et l'indicateur au premier degré pourrait viser une activité qui s'est exercée un ou même deux ans auparavant. On ajoute, bien sûr, aux probabilités de ne pas reconnaître toute l'activité en question dans l'année d'enquête. Il est possible d'envoyer des questionnaires au second degré après avoir traité des réponses au premier degré, et ce, pour la même année que dans la première étape, mais on se retrouve avec un très long décalage entre la fin de l'année de référence et le moment où sont produits les résultats.

Troisièmement, les réponses obtenues au premier degré peuvent être entachées d'une erreur de mesure. Beaucoup de caractéristiques rares dans la population seront aussi d'une définition complexe et, même avec des notes complètes et un questionnaire bien conçu, les entreprises pourraient confondre les choses et mal répondre. Bien entendu, certains ne liront pas les notes, ce qui accroît d'autant les risques de mauvaise réponse. Les études d'erreurs de mesure sont coûteuses. Nous n'avons pas réuni directement de données sur la question, mais les estimations peuvent se révéler très sensibles à un tel traitement d'enquête, surtout si la population a pu être mal estimée au premier degré.

3.3 Note sur l'échantillonnage aux fins des enquêtes sur les marchandises

Les enquêtes sur les marchandises peuvent être considérées comme offrant une version de l'échantillonnage de caractéristiques rares où chaque membre de la population cible présente une telle caractéristique, mais où celle-ci varie selon les unités. Dans ce cas, il est possible d'exploiter un registre satellite d'entreprises écoulant des produits rares. De fait, l'ONS compte un petit nombre de petites enquêtes – qu'il mène principalement pour d'autres ministères – où il adopte cette optique. Un exemple en est son enquête sur les matériaux de construction.

Plus généralement cependant, l'information sur les marchandises est utilisée dans un échantillon à deux degrés où on vise des caractéristiques bien précises. Il y a, par exemple, l'indice des prix à la production (IPP) avec son échantillon à deux degrés fondé sur les données sur les marchandises recueillies dans le cadre de l'enquête PRODCOM. Pour l'IPP, il faut une taille d'échantillon acceptable pour chacun des éléments d'une longue liste de marchandises. Nous pouvons mieux prévoir de bonnes tailles d'échantillon à cause de l'information auxiliaire suffisante que nous donnent les classifications, mais seule l'enquête spécifique sur les produits nous permet de reconnaître les produits qui sont effectivement réalisés par les entreprises. C'est donc là une simple illustration de la même structure d'enquête à deux degrés, mais les questions filtres sont plus pointues, car on a besoin sur toutes les marchandises de données très fines qui servent ensuite à une enquête de seconde étape relativement moins coûteuse (elles servent aussi à établir les estimations de production, bien sûr).

4. ÉCHANTILLON PERMANENT TIRÉ DE DONNÉES ANTÉRIEURES

4.1 Codage de caractéristiques rares

Si les entreprises présentant des caractéristiques rares doivent former une population, il n'y aura ordinairement aucune source qui nous permette de les distinguer des autres entreprises de l'IDBR et, par conséquent, il faudra des mécanismes pour faire la collecte de l'information sur la caractéristique rare si elle est constatée, pour faire le stockage et la mise à jour de cette information. Pour citer un exemple, l'ONS produit des données sur le commerce international de services auquel se livrent une faible minorité d'entreprises de l'industrie tertiaire. De ce type d'enquêtes, on dit généralement à l'ONS qu'il s'agit de « recherches d'aiguilles dans des bottes de foin ».

On mène habituellement une première activité consistant à réunir l'information sur laquelle reposera une enquête du genre, souvent à l'aide d'une question filtre dans une enquête menée auprès d'une population plus vaste (voir plus loin). Par cette question, on reconnaît les entreprises possédant la caractéristique rare d'intérêt et porte un code dans leurs entrées respectives au registre. Les entreprises ainsi codées forment alors une strate à tirage complet dans le

plan de sondage qui est appliqué par la suite. Il est toutefois improbable qu'on repère de la sorte toutes les entreprises d'intérêt, surtout si on considère qu'il n'y a pas de recensement des entreprises à l'ONS. On ne parvient jamais à un dénombrement complet sauf si la caractéristique en question se limite aux grandes entreprises comprises dans les strates à tirage complet. Même là, la non-réponse sera un facteur qui limitera la couverture complète des entreprises. Pour produire une estimation pour les autres entreprises exerçant la même activité d'intérêt (et subsidiairement pour la caractérisation de ces entreprises), on doit faire un certain échantillonnage dans le reste de la population. Cet échantillonnage restera cependant plutôt clairsemé d'ordinaire à cause des restrictions de ressources et, par conséquent, les estimations d'activité seront particulièrement susceptibles d'accuser une grande variabilité dans ce volet du plan de sondage. Il demeure que, avec un échantillon qui embrasse toute la population d'entreprises, on a l'avantage de pouvoir apporter des corrections en fonction de l'erreur de mesure (par la question filtre) décelée dans un échantillonnage à deux degrés (voir la section 3.2), puisque chaque membre de la population aura des probabilités non nulles d'être inclus dans l'échantillon. Sans pousser l'analyse, on ne saura au juste si l'erreur quadratique moyenne (eqm) est plus petite sans erreur de mesure – mais avec une grande variabilité – ou si nous pouvons accepter une certaine erreur de mesure dans cette réduction.

Dans d'autres cas, il y aura peut-être une source par laquelle nous pourrons créer le codage de caractéristiques rares : presse (les revues de l'industrie du bâtiment servent à diverses enquêtes portant sur les matériaux de construction, par exemple), listes de Dun & Bradstreet, pages jaunes, etc.

4.2 Mise à jour du codage

On tient à jour le codage d'activités rares en retranchant les entreprises qui ont cessé d'exercer l'activité en question et en ajoutant celles qui viennent tout juste de s'y adonner. Les entreprises qui mettent fin à une activité aussi particulière sont relativement rares. D'habitude, on les conserve un certain temps dans la partie fortement échantillonnée de la population (selon une « règle d'inertie ») et leur enlève leur code seulement s'il devient clair qu'elles ont définitivement cessé d'exercer l'activité d'intérêt. C'est la bonne chose à faire là où le coût de réidentification (en cas de reprise de l'activité) est élevé par rapport au coût (questionnaires envoyés pour rien et variances amplifiées) de leur maintien dans l'enquête même après cessation de l'activité. Il y a régulièrement des entreprises nouvellement reconnues qui viennent de l'échantillonnage du reste de la population; elles sont appelées à passer de la partie faiblement échantillonnée à la partie fortement échantillonnée de la base de sondage.

La mise à jour du codage risque de créer d'importantes différences dans les estimations d'activité entre les années. Dans des enquêtes qui se succèdent, une entreprise au niveau d'activité relativement élevé pourrait ne plus être échantillonnée à chaque coup (et donc représentée principalement par les autres entreprises n'ayant pas l'activité en question) et sélectionnée tour à tour dans la partie faiblement échantillonnée (son activité recevra alors un poids important en raison de cet échantillonnage limité) et dans la partie fortement échantillonnée (son activité recevra alors un poids beaucoup plus petit et le total sera bien plus bas). C'est en fait ce qui se produit dans les données du commerce international des services. Le phénomène était particulièrement perceptible lorsqu'on a étendu l'échantillonnage à une partie de la population d'entreprises qui était censée ne pas exercer l'activité rare en question. La première année, il y a eu beaucoup d'activité nouvelle et des valeurs importantes de pondération, puis une baisse des valeurs lorsque les entreprises en question sont passées à la partie fortement échantillonnée. Cette baisse est due à la variabilité d'échantillonnage (elle est l'expression de la grande variance caractéristique de ce type de plan de sondage), bien que, dans le cas de la première année et de l'année qui suit, on puisse avoir raisonnablement considéré que le plan de sondage avait changé et que la variabilité d'échantillonnage changeait de ce fait.

4.3 Estimations cohérentes des variations

Plusieurs possibilités s'offrent pour l'estimation des variations en cas de grande variabilité entre périodes. Il y a d'abord une stratégie de traitement des valeurs aberrantes permettant de réduire l'importance relative de la poignée d'entreprises ayant beaucoup d'activité dans la partie faiblement échantillonnée lorsqu'on établit l'estimation relative aux entreprises non échantillonnées. C'est le problème classique des valeurs extrêmes dans les enquêtes par sondage (voir, par exemple, Hulliger, 1995), mais il est beaucoup plus important ici à cause de la nature des réponses (oui/non). Karlberg (2000) a conçu et éprouvé des méthodes d'estimation des totaux pour des populations

où les réponses néant sont nombreuses en proportion, mais on ne les a pas appliquées jusqu'à présent à l'estimation des variations dans les enquêtes de l'ONS que nous avons décrites. Il est mois clair si elles sont d'une stabilité suffisante pour une bonne estimation des variations entre années.

Une autre stratégie consiste à lisser dans le temps. Il s'agit principalement ici d'en venir à une estimation du nombre de « valeurs aberrantes » (réponses positives) dans la population en regroupant les données pour un certain nombre de périodes, ce qui permet de fournir des totaux de population pour la poststratification et donc d'accroître la précision, au prix cependant d'une certaine perte de réponse si la proportion d'entreprises exerçant l'activité rare en question varie d'année en année.

4.4 Exclusions

Dans certaines parties de la base de sondage, l'activité rare peut être si clairsemée qu'il devient par trop coûteux de l'inclure dans un plan d'échantillonnage efficient. Nous pourrions alors décider d'exclure cette partie de l'échantillonnage. C'est en fait ce que nous faisons pour les plus petites entreprises (comptant moins de 10 salariés) dans le volet « commerce international de services », leur fréquence étant si faible. Nous avons alors besoin d'une information suffisante pour des entreprises « assimilées » (qui seront ordinairement un peu plus grandes) avec peut-être une variable auxiliaire pour nous aider à construire un modèle d'estimation. Il y aura aussi la possibilité pour nous de supposer que l'activité en question n'est tout simplement pas là. Dans l'un et l'autre cas, nous acceptons un certain biais (de modèle) pour exclure la variance due à l'échantillonnage et garder les coûts à un niveau acceptable.

5. EXAMEN

Dans les enquêtes qui sont des « recherches d'aiguilles dans des bottes de foin », il est difficile de bien couvrir la population visée si on ne dispose pas d'une source administrative, mais même si on peut trouver des sources appropriées, on aura différents défis à relever. Souvent, on aura un choix de sources incomplètes, auquel cas on devra s'efforcer de les harmoniser. Parfois aussi, les sources seront tout à fait indépendantes les unes des autres et se situeront à des niveaux différents. Dans un cas extrême comme celui de la différence évoquée à la section 2 entre les caisses et les régimes de retraite, il sera peut-être bon de songer à un échantillonnage en réseau pour pouvoir relier caisses et régimes dans l'enquête.

Les enquêtes à deux degrés relèvent d'une théorie bien définie, mais elles sont relativement lentes. Il y a des façons d'accroître leur rendement, ce qui nous mènera à un type de panel post-stratifié comme ceux que nous avons décrits à la section 4 avec une variance réduite dans les strates à tirage complet et sans biais d'erreur de mesure malgré le risque de grande variance dans l'ensemble en raison de la rareté des observations.

Toutes ces orientations ont certaines caractéristiques en commun qui les distinguent de toute autre enquête auprès des entreprises. Dans ces enquêtes, on dispose normalement de peu de variables auxiliaires, mais la situation est pire dans bien des cas faute de sources indépendantes d'information sur des caractéristiques rares. Même s'il était possible d'obtenir un certain nombre de variables auxiliaires par jumelage d'un registre satellite et d'un registre principal, elles seront souvent d'une piètre capacité prévisionnelle pour les grandes variables d'intérêt et n'aideront pas à réduire la variabilité (dans certaines des enquêtes financières de l'ONS, la différence entre les estimations de Horvitz-Thompson et les estimations par le quotient réside dans la normalisation, puisqu'il y a peu de différence de variances).

Que les estimations de niveaux soient les meilleures possible n'est habituellement pas le souci premier dans ces enquêtes dont on se sert pour l'observation des conditions et l'élaboration de politiques dans le domaine économique. Ce sont souvent les mouvements ou variations d'une série qui auront le plus d'intérêt. Avec tous les problèmes d'estimation de niveaux, il est difficile de mettre au point une méthode qui sera approximativement sans biais et se caractérisera aussi par de faibles variances. En dernière analyse, la seule façon de produire des estimations raisonnables de variations est de s'assurer que, « pour la plupart », les unités qui exercent l'activité rare visée se retrouveront dans l'échantillon. Même là, on aura le défi de faire connaître les lacunes des estimations aux utilisateurs.

RÉFÉRENCES

Estevao, V. M. et Särndal C.-E. (2002), "The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling", *Journal of Official Statistics*, 18, pp. 233-255.

HMCE (2002), "Notice 700/17 – Funded Pension Schemes", www.hmce.gov.uk

Hulliger, B. (1995), "Estimateurs Horvitz-Thompson à l'épreuve des valeurs aberrantes", *Techniques d'enquête*, 21, pp. 89-97.

Karlberg, F. (2000), "Survey estimation for highly skewed populations in the presence of zeroes", *Journal of Official Statistics*, 16, pp. 229-241.

Perry, J. (1995), "The Inter-Departmental Business Register", *Economic Trends*, 505, Novembre, pp. 27-30.