LA QUALITÉ DES ESTIMATIONS DE LA POPULATION ACTIVE TIRÉES DE LA CURRENT POPULATION SURVEY : PERSPECTIVES D'ENQUÊTE GLOBALE

Lawrence Cahoon, Patrick Flanagan et Karen Deaver[1]

RÉSUMÉ

L'enquête américaine sur l'état de la population est la principale source de données sur la population active aux États-Unis. Il est impératif que la qualité des données soit maintenue tout au long du processus d'enquête. Les auteurs décrivent comment la qualité est abordée durant ce processus, depuis la base de sondage jusqu'aux estimations, en passant par toutes les étapes intermédiaires. Ils traitent aussi de l'élaboration de la base de sondage, des opérations d'échantillonnage, du contrôle de l'échantillon, de la collecte des données, de la vérification, de l'imputation, de l'estimation, de l'élaboration du questionnaire et des évaluations de la qualité incorporées dans le processus d'enquête. L'analyse se termine par une discussion des recherches en cours et des améliorations susceptibles d'être apportées à l'enquête.

MOTS CLÉS : Qualité des données, erreur d'échantillonnage, erreur non liée à l'échantillonnage, contrôle de la qualité

1. INTRODUCTION

1.1 La Current Population Survey

Aux États-Unis, l'enquête sur l'état de la population (soit la * Current Population Survey + ou la * CPS +) constitue la principale source de statistiques mensuelles sur la situation d'activité et les caractéristiques démographiques des travailleurs. Les responsables de certains programmes socio-économiques d'envergure nationale s'inspirent notamment des statistiques produites par la CPS pour apporter des changements d'orientation et pour mieux répartir les dépenses fédérales. Par ailleurs, des études réalisées par des chercheurs du monde entier se servent des données de la CPS pour valider toutes sortes d'hypothèses. Vu l'énorme responsabilité rattachée à l'exactitude des données et des estimations statistiques, tous les aspects de la qualité des données revêtent une importance accrue.

1.2 Historique de la CPS

Les origines de la CPS remontent à 1940. Intitulée à l'origine * Sample Survey of Unemployment +, l'enquête a été lancée par la Work Projects Administration en mars 1940. Elle a été transférée au Bureau of the Census en août 1942 et renommée * Monthly Report on the Labor Force + en 1943. Quatre ans plus tard, elle a pris sa

[1]Lawrence Cahoon, Patrick Flanagan et Karen Deaver, U. S. Bureau of the Census, Washington, DC, 20233, USA

dénomination actuelle. Depuis, on en a constamment amélioré le plan de sondage et les méthodes d'enquête, à la faveur des dernières percées méthodologiques et des résultats des innombrables études sur la qualité des données de la CPS réalisées par des chercheurs. Le lecteur trouvera dans Bureau of the Census (2000) l'évolution détaillée du plan de la CPS de ses origines jusqu'à présent, notamment la refonte décennale de l'échantillon.

1.3 Refonte décennale de la CPS

Dans la foulée du recensement décennal des États-Unis, la CPS et d'autres grandes enquêtes démographiques font appel à l'information mise à jour pour actualiser leur plan et sélectionner de nouveaux échantillons en vue des dix prochaines années.

2. CONTRÔLE DE LA QUALITÉ DES DONNÉES DE LA CPS

2.1 Plan de l'enquête

Le plan de sondage de la CPS met surtout l'accent sur l'estimation de l'emploi parmi la population civile (c'est-à-dire non militaire) hors établissement âgée de 16 ans et plus. Plus précisément, il met en évidence le taux de chômage. Les objectifs de ce plan sont les suivants :

- Atteindre un coefficient de variation(CV)[2] de 1,9 % en ce qui concerne le niveau mensuel de chômage, dans l'hypothèse d'un taux de chômage de 6 %. Cet objectif est fondé sur la nécessité de relever une variation de 0,2 % du taux de chômage d'un mois à l'autre.
- Atteindre un CV maximum de 8 % en ce qui concerne l'estimation du niveau annuel de chômage pour chaque État, dans l'hypothèse d'un taux de chômage de 6 %.

On atteindrait sans peine les objectifs ci-dessus si on disposait d'un budget illimité, mais avec des si... Le plan actuel de la CPS fait appel à l'échantillonnage en grappes afin d'accroître l'efficience et il réussit à atteindre les objectifs avec un échantillon de 50 000 unités de logement par mois renfermant environ 112 000 personnes.

La sélection de l'échantillon se fait en deux temps. Premièrement, on choisit les unités primaires d'échantillonnage (UPE) géographiques et deuxièmement, les unités de logement.

Chaque ménage sélectionné fait partie de l'échantillon pendant quatre mois consécutifs, en est retiré pendant huit mois, puis est réintégré pour quatre mois consécutifs.

2.2 Sélection de l'échantillon

2.2.1 Sélection de l'échantillon - Premier volet

2.2.1.1 Vue d'ensemble

Les États-Unis sont répartis en 2 007 UPE, consistant chacune en un ou plusieurs comtés. Parmi les principales UPE de la population, un bon nombre sont sélectionnées avec certitude : ce sont les UPE autoreprésentatives (AR). D'ailleurs, les UPE qui font partie des 150 plus grandes régions statistiques métropolitaines sont automatiquement considérées comme AR. Les UPE restantes, dites non autoreprésentatives (NAR), sont

[2] Le coefficient de variation (CV), également désigné * écart-type relatif +, représente l'écart-type de l'estimation divisé par sa valeur attendue.

constituées en strates au sein de l'État où elles se trouvent, et une UPE est sélectionnée à partir de chaque strate.

On fait aussi appel à un algorithme additionnel au moment de sélectionner les UPE NAR. Cet algorithme accroît la probabilité qu'une UPE sélectionnée dans le plan de 1990 sera à nouveau sélectionnée au moment de la refonte de 2000, tout en maintenant fixe la probabilité globale de sélection selon une probabilité proportionnelle à la taille. L'utilisation de cet algorithme a une incidence sur la qualité des données – et plus précisément sur la variance des estimations – du fait qu'elle crée une dépendance entre les strates d'un même État. Aucune étude n'a été publiée à ce jour concernant l'ampleur de cette incidence.

La sélection des UPE comporte quatre volets. Elle se fait à partir des fichiers officiels du recensement décennal, assujettis à un processus de vérification qui leur est propre.

2.2.1.2 Élaboration du fichier des définitions de l'UPE

On élabore un fichier précisant, à côté de chaque numéro d'UPE, le comté correspondant. Les UPE sont définies selon des critères géographiques particuliers (énumérés dans Bureau of the Census (2000)) portant sur les contraintes et les charges de travail en matière d'interview. Les définitions des UPE sont créées à partir des UPE de la décennie précédente. Après avoir mis à jour les définitions des comtés, nous sollicitons l'apport de notre personnel des bureaux régionaux pour ce qui concerne les problèmes posés par les définitions antérieures. Les nouvelles définitions prennent également en compte les projections concernant les nouvelles régions statistiques métropolitaines qui doivent être considérées comme AR.

Contrôle de la qualité : Une fois le fichier créé, il est réparti parmi les membres d'un groupe de travail par État. Les vérificateurs se penchent sur chaque UPE de leur État pour s'assurer qu'elle répond systématiquement à la définition et qu'elle incorpore les recommandations du personnel sur place qui ont été retenues. Les vérificateurs font aussi appel à des cartes afin de comparer les nouvelles définitions d'UPE aux anciennes et de s'assurer que tous les comtés sont pris en compte.

2.2.1.3 Élaboration du fichier principal de stratification

Lorsque le fichier des définitions d'UPE et le fichier d'entrée du recensement décennal sont à point, l'étape suivante consiste à construire le fichier principal de stratification, établi par comté et comprenant la définition du comté, un numéro d'UPE, la taille de la population âgée de 16 ans et plus au sein du comté, le nombre d'unités de logement et les variables de stratification sélectionnées. Celles-ci sont étroitement corrélées avec le chômage. À partir de ce fichier initial, on établit un fichier composite au niveau de l'UPE pour les besoins du processus de stratification. Au moment de construire le fichier, certaines variables conservent la forme qu'elles avaient dans le fichier d'entrée du recensement, alors que d'autres sont calculées à partir de variables d'entrée (par exemple, les données sommaires sur le comté).

Contrôle de la qualité : Les spécifications logicielles et le logiciel de dépouillement sont soumis à un essai de système exhaustif avant la production en vue de vérifier les algorithmes et de garantir la présence de toutes les variables. Les fichiers d'entrée sont également vérifiés à l'aide de programmes de vérification. Un ensemble de production initial portant sur trois États fait l'objet d'une vérification systématique pour garantir la présence de chaque variable requise et la conformité des fourchettes de valeurs aux normes; il est ensuite comparé aux fichiers de refonte précédents pour faire ressortir les observations aberrantes. Les États restants sont vérifiés au moyen de calculs sommaires.

2.2.1.4 Stratification des UPE

Le processus de stratification fait appel à un programme de recherche fondé sur l'estimation de la variance à partir de l'algorithme de Friedman-Rubin. Au sein de chaque État, ce programme constitue des strates aléatoires basées sur l'information concernant les restrictions sur la taille et l'homogénéité des variables de stratification. Il calcule ensuite les estimations de la variance pour les principales variables de chaque État et pour le pays dans son ensemble. Puis, il utilise un processus itératif, échangeant les UPE de diverses strates dans chaque État jusqu'à l'obtention d'une stratification optimale. Le programme de recherche tient également compte de la mise en équilibre de la charge de travail à partir des objectifs fixés.

Contrôle de la qualité : Ce logiciel s'inspire de celui fondé sur l'algorithme de Friedman-Rubin, utilisé pour les besoins de la stratification de 1990. La principale mesure de contrôle de la qualité (CQ) a été la confirmation des algorithmes utilisés, conjuguée à la révision structurée du logiciel pour confirmer que celui-ci exécutait les algorithmes comme prévu. La plus grande partie de ce CQ a été confiée à un groupe de travail composé de statisticiens-mathématiciens, dont les résultats ont été soumis à l'examen de notre division de la recherche et du développement pour fins de confirmation de l'exactitude technique. Enfin, le programme a fait l'objet d'essais exhaustifs utilisant des données réelles simulées fondées sur les fichiers de données de recherche de 1990. La vérification en phase de production comprendra une comparaison des résultats entre la dernière décennie et la refonte en cours dans chaque État et portera sur les définitions des strates, le nombre de strates et l'estimation des variances entre les UPE.

2.2.1.5 Sélection des UPE

La sélection des UPE comporte en fait deux processus. Le premier aborde chaque strate NAR séparément et calcule les probabilités de sélection qui serviront au second processus, celui de la sélection proprement dite. Le premier processus calcule la probabilité * globale + de sélection pour chaque UPE en fonction d'une probabilité proportionnelle à la taille, laquelle correspond à la population de l'UPE âgée de 16 ans et plus. Il s'agit ensuite d'ajouter l'information sur la sélection tirée du plan de 1990. On est alors en mesure de calculer les probabilités conditionnelles de sélection pour chaque UPE, à l'aide d'un algorithme qui traite les sélections de 1990 et de 2000 comme une seule expérience. Après avoir tenu constantes les probabilités globales de sélection de chaque UPE, cet algorithme calcule les probabilités conditionnelles de sélection pour les UPE de la refonte de 2000 compte tenu des UPE sélectionnées dans le plan de 1990. Cette maximisation du chevauchement réduit considérablement le nombre d'UPE qui changement d'un plan à l'autre, atténuant de ce fait l'altération de la qualité des données qui se produirait si le Bureau of the Census devait engager de nouveaux intervieweurs dans les nouvelles UPE et remercier des intervieweurs dans les anciennes. Le lecteur trouvera dans Bureau of the Census (2000) plus de détails sur l'algorithme.

Contrôle de la qualité : Le contrôle initial de la qualité de ce processus consiste en la vérification de l'exactitude statistique des algorithmes à utiliser. La méthodologie a été définie dans un document de planification qui a été passé en revue par les statisticiens du groupe de travail chargé de la mise au point, puis par notre division de la recherche et du développement pour en confirmer l'exactitude statistique. On a ensuite rédigé une spécification précisant la méthode retenue pour incorporer les algorithmes dans le logiciel, et on en a vérifié la concordance avec le plan de méthodologie. Le logiciel résultant a ensuite fait l'objet d'essais exhaustifs sur une combinaison de données réelles et de données simulées comparant les résultats de tous les calculs avec des résultats indépendants. Enfin, on a vérifié le dernier passage de production pour s'assurer qu'une vaste gamme de calculs sommaires soient conformes à des fourchettes et à des totaux prédéterminés.

2.2.2 Sélection de l'échantillon – Deuxième volet

Une fois les UPE sélectionnées, il faut sélectionner les unités de logement au sein de chacune. La liste des unités de logement tirée du recensement représente la principale source de renseignements à cet égard. Pour les besoins du recensement de 2000, cette liste est désignée le fichier maître des adresses (FMA), mais elle

comporte des lacunes pour la CPS.

2.2.2.1 Problèmes liés à l'utilisation du FMA décennal comme unique base de sondage

Le recensement s'est surtout fait par voie postale. Dans certaines régions, les adresses renvoient à une case postale plutôt qu'à une unité de logement. De plus, bien des unités de logement figurant sur la liste comportent uniquement une description physique. De plus, il faudra mettre à jour la liste du FMA tout au long de la décennie pour tenir compte des changements de logement. Pour régler ces problèmes, nous avons fait appel à quatre bases pour la sélection de l'échantillon.

2.2.2.2 Couverture par quatre bases

Dans les îlots où le FMA comporte au moins 96 % de * bonnes + adresses, nous utilisons le FMA comme base. Pour les unités de logement, il s'agit de la **base unitaire.** Pour les logements collectifs comme les résidences universitaires, il s'agit plutôt de la **base des logements collectifs.** Pour les autres îlots, nous créons une **base aréolaire,** au sein de laquelle nous sélectionnons des unités de logement et des logements collectifs pour l'échantillon à partir d'îlots et de dénombrements d'îlots; l'îlot fait ensuite l'objet d'un listage en vue de repérer les unités de logement effectivement incorporées dans l'échantillon. Pour tenir compte des logements nouvellement construits, nous disposons d'une **base de permis** dans les régions du pays où les permis de construire sont émis. Dans les autres régions, nous nous fions à la base aréolaire pour mettre en relief les nouvelles constructions.

2.2.2.3 Lacunes de couverture de la base

La lacune de couverture la plus évidente tient aux erreurs que renferme la base unitaire tirée du recensement. En effet, les logements absents de la liste du recensement, mais compris dans la base unitaire, sont exclus. Il s'agit d'un facteur de couverture négligeable, car le recensement fait appel à un dénombrement par îlots, et les îlots faisant partie de la base unitaire se trouvent habituellement dans les régions où il est le plus facile de trouver une adresse. Un deuxième facteur de couverture, soit le décalage lié aux permis, résulte des permis émis avant le jour du recensement pour des logements construits ultérieurement. La base des permis de la CPS tient compte des permis remontant au mois de janvier précédant le jour du recensement, en vue de régler à la fois ce problème de décalage et l'éventuel problème de surdénombrement résultant de la double sélection des nouveaux immeubles. Un troisième problème a trait à la couverture des maisons mobiles. Dans les régions visées par une base unitaire, aucun système ne permet à ce jour d'actualiser les listes de la base afin de prendre en compte les nouvelles maisons mobiles qui s'y installent.

Une fois terminée la sélection des UPE, il faut procéder à la sélection des unités de logement de chaque base. L'élaboration de la base et la sélection des unités de logement comportent trois grands processus, dont chacun est assorti de tests du système de mise au point du programme informatique et de la vérification détaillée de la production pour en garantir l'exécution adéquate. L'information nécessaire à cette étape est tirée du FMA du recensement décennal, lequel a été soumis à son propre processus de vérification, et des listes du bureau d'octroi des permis provenant des fichiers de l'enquête sur la construction.

2.2.2.4 Tri servant à la définition de la base

Dans les UPE sélectionnées, il faut soumettre chaque îlot à un tri afin de déterminer les îlots qui comptent suffisamment de * bonnes + adresses et ceux qui n'en comptent pas. Les îlots ayant les * bonnes + adresses sont affectés à la base unitaire et à celle des logements collectifs (voir le paragraphe 2.2.2.2) et les autres, à la base aréolaire. Le tri se fait automatiquement au moyen d'un programme informatique.

Un deuxième tri permet de déterminer les îlots au sein des UPE sélectionnées qui sont visés par un bureau d'octroi des permis et ceux qui ne le sont pas. Les îlots ayant de * bonnes + adresses, mais non visés par un

bureau d'octroi, sont transférés de la base unitaire à la base aréolaire.

L'affectation des adresses à la base des logements collectifs est relativement simple, du fait que les fichiers du recensement décennal font ressortir les logements collectifs et nous permettent de ce fait de séparer ces adresses de la base unitaire et de mettre à jour notre liste de l'univers des logements collectifs.

Contrôle de la qualité : Les critères servant à définir les * bonnes + adresses sont spécifiés et passés en revue par tous les groupes intéressés au moment de la mise au point du logiciel de tri. Une fois ce logiciel terminé, un échantillon d'adresses réelles est trié par le programme et par un logiciel indépendant pour confirmer le bon fonctionnement du processus.

2.2.2.5 Création de la base

À cette étape, il s'agit principalement de constituer les fichiers aux fins de la sélection. La base unitaire et celle des logements collectifs sont, pour l'essentiel, constituées à partir des fichiers du recensement décennal, même si quelques données agrégées du questionnaire détaillé du recensement sont ajoutées pour les besoins des tris servant à la sélection de l'échantillon. S'agissant de la base aréolaire, les fichiers sont constitués par îlots. Bien que les adresses ne soient pas utiles lorsqu'on procède à la sélection, les chiffres du recensement par îlot servent à constituer une structure de sélection qui renfermera les adresses spécifiques au moment du listage. Enfin, la base des permis de construire établit une base * squelette + fondée sur la croissance prévue au niveau des îlots.

Contrôle de la qualité : Parallèlement aux autres mesures de contrôle de la qualité, le traitement est vérifié à partir de fichiers d'essai de données réelles du recensement et d'estimations au niveau des îlots. Les étapes de ce processus sont contrôlées par une combinaison de comparaisons avec des logiciels indépendants et de calculs sommaires.

2.2.2.6 Sélection de l'échantillon

Pour chaque base et au sein de chaque UPE, les unités de logement sont regroupées en unités d'échantillonnage ultimes comportant chacune quatre unités, afin de tirer parti de l'efficience de l'échantillonnage en grappes. Ces unités sont ensuite triées selon une série de variables, dont l'urbanisation, la race, le mode d'occupation et le sexe. Enfin, un échantillon systématique est tiré par la sélection d'une série de 21 unités d'échantillonnage ultimes à chaque point de sélection. Il en résulte 21 échantillons qui doivent servir pendant toute la décennie.

Contrôle de la qualité : Un échantillon de données portant sur six États est soumis au processus. Chaque étape est vérifiée à la main ou par une comparaison fondée sur un logiciel indépendant. Les statistiques sommaires sont vérifiées à cette étape et au moment de la production.

2.2.2.7 CQ de la liste de la base aréolaire

Le contrôle de la qualité du listage aréolaire se fait à la fois par la formation initiale des responsables du listage et par la vérification de la qualité du listage par chaque préposé. Les préposés au listage font l'objet d'une vérification annuelle, au cours de laquelle un agent principal sur place soumet un échantillon d'îlots ou de segments à un listage de contrôle. Lorsqu'un préposé a un taux d'erreur inacceptable, il retourne en formation ou est affecté à d'autres tâches.

2.3 Collecte des données

2.3.1 Le questionnaire

La dernière révision du questionnaire de la CPS remonte à 1994, année de la conversion à l'automatisation.

La formulation des questions, la terminologie et l'enchaînement des questions évoluent depuis des années, à la faveur d'études exhaustives sur l'exactitude des données résultantes, de tests cognitifs et d'analyses de l'erreur de réponse. En outre, la conversion aux interviews sur place assistées par ordinateur (IPAO) et aux interviews téléphoniques assistées par ordinateur (ITAO) a atténué davantage les problèmes qu'ont les interviewers à suivre les instructions * passez à + de l'enquête.

2.3.2 Méthode d'interview

Pendant le premier mois où le logement est dans l'échantillon, les interviews se font pour l'essentiel (80 %) sur place afin d'établir l'identité des membres du ménage et de recueillir des renseignements sur le ménage. Par la suite, la plupart des interviews (85 %) se font par téléphone, sauf au cours du cinquième mois, où à peu près 65 % des interviews sont faites sur place.

2.3.3 Processus d'interview

Pour les besoins de la CPS, les interviews prennent surtout la forme d'IPAO confiées à des représentants sur place par l'entremise des bureaux régionaux. Par contre, certaines cas (environ 10 %) sont des ITAO confiées à l'un des trois centres téléphoniques. Le nombre d'interviews confiées à chaque centre est déterminé par un processus d'équilibrage de la charge de travail. En dépit de ce que leur nom peut donner à entendre, les IPAO peuvent se faire en personne ou par téléphone. Nous utilisons en effet le terme IPAO lorsque le représentant se sert d'un ordinateur portable pour saisir les données sur place. Qu'on fasse appel à l'IPAO ou à l'ITAO, le questionnaire amène automatiquement l'interviewer à travers chaque étape de l'enquête, ce qui lui permet de saisir les données à l'endroit voulu.

2.3.4 Réponse "par publicayion"

Le questionnaire de la CPS s'intéresse d'abord au ménage dans son ensemble, puis à chaque membre du ménage. Si chacun des membres est présent, il peut répondre directement aux questions. Sinon, la CPS permet à un membre du ménage (âgé de 15 ans ou plus) de répondre pour les autres. En règle générale, des réponses par procuration sont obtenues répondent pour environ 50 % de toutes les personnes dans l'échantillon.

2.3.5 Influence de l'intervieweur

La variation de la qualité des données d'un intervieweur à un autre représente un phénomène bien documenté (Lyberg et Kasprzyk, 1991). Il ne s'agit pas uniquement du niveau de réponse atteint par l'intervieweur. Certains intervieweurs réussissent tout simplement mieux à établir un lien avec les répondants. Ce facteur exerce une influence plus profonde sur la CPS, une enquête par panel, que sur les enquêtes ponctuelles ou transversales. L'intervieweur constitue en effet une source de biais du fait que les intervieweurs ne répondent pas tous de la même façon aux questions des enquêtés et qu'ils ne suscitent pas les mêmes réactions de leur part. Par exemple, si l'enquêté se sent mal parce qu'il a perdu son emploi et qu'il éprouve un certain respect à l'égard de l'intervieweur, il risque de donner une réponse inexacte à la question sur l'emploi.

La formation des intervieweurs pour les besoins de la CPS se fait habituellement en deux étapes, en supposant que l'intervieweur a déjà été initié aux rudiments de l'interview. Dans un premier temps, l'intervieweur reçoit une formation initiale consistant en études à domicile, en formation sur place et en deux jours d'observation sur le tas. Au cours du deuxième mois, le nouvel intervieweur doit à nouveau étudier à la maison et se prêter à une journée d'observation. Enfin, la formation initiale est complétée par l'étude à domicile et un dernier examen de révision avant le troisième mois. Par la suite, l'intervieweur fait mensuellement des études à domicile et il assiste deux fois l'an à un cours de perfectionnement.

2.3.6 Erreurs dues à la liste

Il y a sous-dénombrement lorsque la liste établie au moment de l'interview initiale n'englobe pas tous les membres du ménage. D'après une étude réalisée en marge du recensement décennal, à peu près 32 % de toutes les personnes omises appartenaient à un ménage dont la totalité des membres ne figuraient pas sur la liste du ménage (Killion, 1993).

2.3.7 Réinterview pour les besoins du CQ

Chaque mois, un échantillon des représentants sur place affectés à la CPS est sélectionné, avec un échantillon des cas qui leur sont confiés. Les enquêtés en question sont contactés une deuxième fois afin de confirmer que le représentant a mené l'interview dans le respect des normes établies. Lorsqu'on relève une erreur significative ou qu'on soupçonne l'intervieweur d'avoir faussé les résultats, le superviseur de la CPS est saisi du cas et chargé d'y donner suite.

2.3.8 Contrôle des ITAO

Les installations d'ITAO sont dotées d'un programme de contrôle assimilé au processus de réinterview. En vertu de ce programme, un * encadreur + choisit au hasard des interviews et en suit le déroulement. Il s'assure que toutes les étapes sont suivies, puis débreffe l'intervieweur en faisant le point sur son rendement.

2.3.9 Suivi des cas

Tel qu'il est précisé au paragraphe 2.6.1, les cas sont suivis au siège du Bureau of the Census, aux bureaux régionaux et dans les installations d'ITAO afin de garantir qu'aucun cas n'est perdu au moment du transfert d'un fichier.

2.4 Traitement des données

2.4.1 Transmission des données

Une fois complétés par le représentant sur place, les cas sont transmis par modem à la base de données du bureau régional. S'il s'agit d'une interview achevée, elle est regroupée avec les autres interviews achevées et acheminée au siège du Bureau of the Census. Les cas qu'un représentant ne peut achever peuvent être réaffectés au superviseur de la CPS, habituellement à un représentant principal. La responsabilité de tous les cas incombe à la fois aux bureaux régionaux et au siège afin de garantir l'achèvement et le retour de chaque cas.

2.4.2 Codage

Au moment du traitement, un code est attribué à chaque entrée textuelle portant sur le secteur d'activité et la profession de chaque personne dans l'échantillon. Des codeurs spécialement formés sont affectés à cette tâche dans le centre de traitement national du Bureau of the Census.

La qualité du codage du secteur d'activité et de la profession est contrôlée par un processus de recodage. Un échantillon du travail de chaque codeur est sélectionné, puis codé par un deuxième codeur[3]. Lorsqu'un codeur commet trop d'erreurs au cours d'un mois donné, il doit suivre des cours de perfectionnement. En raison des

[3]Il s'agit en fait d'un codeur d'expérience qui s'occupe habituellement des travaux de codage particulièrement difficiles et qui se charge, accessoirement, du contrôle de la qualité.

courts délais établis pour le traitement de la CPS, ce système n'est pas assimilable à un essai d'acceptation de l'ensemble de données. En revanche, grâce à sa stabilité éprouvée, ce système d'information sert à surveiller le système de codage et à en garantir la stabilité.

2.4.3 Vérification et imputation

Le traitement de la CPS donne lieu à une grande diversité de vérifications. Les entrées hors champ, les réponses * ne sait pas + et les refus peuvent être remplacés par des valeurs induites d'autres entrées ou de réponses à des questions semblables posées au cours d'un mois précédent. La non-réponse à une question est imputée au moyen de la méthode dite * hot deck +, selon laquelle le classement et le tri servent à repérer un donneur qui ressemble à la personne échantillonnée sur le plan de l'âge, de la race et du sexe, par exemple. La réponse du donneur à la même question est imputée à titre d'estimation valable. Étant donné le faible pourcentage de valeurs imputées pour une question donnée (souvent <0,5 %), l'incidence de l'imputation sur les données primaires est considérée comme négligeable.

Dès que les systèmes de vérification et de codage sont modifiés, ils sont soumis à un essai de système exhaustif. Par ailleurs, une fois terminée la vérification et, par la suite, l'imputation, on examine les distributions initiales et finales de chaque élément de donnée afin de relever les variations inhabituelles.

2.5 Estimations

2.5.1 Pondération de base

Comme la CPS n'est pas fondée sur un échantillon aléatoire simple, nous utilisons des poids en vue d'obtenir des estimations non biaisées. Le plan d'échantillonnage de base cherche à créer une situation d'autopondération au sein de chaque État, de sorte que chaque unité de logement aura la même probabilité de sélection. La probabilité de sélection pour un État donné – ou plus exactement son contraire, c'est-à-dire l'intervalle d'échantillonnage global de l'État – est choisie dans le but d'atteindre l'objectif fixé pour le CV de l'État.

Le système de pondération est soumis à un essai de système exhaustif à chaque fois qu'il est modifié. Puis, une fois les poids calculés, les distributions de pondération sont comparées à celles des mois précédents pour voir si on a créé des pondérations inhabituelles.

Dans un deuxième temps, les poids-personne résultant du traitement sont comparés chaque mois aux résultats d'un programme de pondération indépendant. L'écart absolu cumulatif de ces poids doit être inférieur à un seuil fixé d'avance pour que la pondération soit acceptée.

2.5.2 Estimation de la variance

Les variances sont estimées pour la CPS à partir de la méthode des * différences successives + abordée dans Fay et Train (1995), au moyen de 160 répétitions ou réplicats. Le plan ou la programmation de l'estimation de la variance est soumis à un essai de système exhaustif à chaque fois qu'il est modifié.

3. RÉDUCTION DES ERREURS DE LA CPS

3.1 Rajustement de la pondération

Étant donné que certaines formes d'erreur non liée à l'échantillonnage – et notamment la non-réponse – sont inévitables, nous apportons certains rajustements aux poids afin d'atténuer l'incidence que les poids peuvent avoir sur les estimations, au chapitre du biais et de la variance.

3.1.1 Rajustement de la non-réponse

Dans la CPS, nous procédons à un rajustement simple de la non-réponse au sein des cellules selon l'hypothèse voulant que la non-réponse au sein des cellules en question soit répartie aléatoirement par rapport au chômage. Les cellules sont déterminées par la région statistique métropolitaine où se trouve chaque UPE et par la taille de cette région. Le poids des ménages non répondants (type A seulement) est redistribué parmi les ménages répondants de la même cellule.

3.1.2 Rajustement du ratio à base de * contrôles +

La CPS rajuste le ratio des pondérations en deux temps. Dans un premier temps, les poids sont rajustés pour tous les cas dans chaque UPE NAR sélectionnée en fonction du déséquilibre éventuel de la représentation raciale occasionné par la sélection de l'UPE. Dans un deuxième temps, tous les poids des cas sont rajustés selon une méthode itérative du quotient qui fait appel à un contrôle pour la population de l'État âgée de 16 ans et plus, les groupes d'âge/de sexe/d'hispanophones et les groupes d'âge/de sexe/de race itérativement sur six itérations. Les * contrôles + utilisés pour ces rajustements sont des estimations indépendantes de la population civile non institutionnalisée âgée de 16 ans et plus tirées du recensement décennal, mais rajustées d'abord pour tenir compte du sous-dénombrement et ensuite au moyen de données sur les naissances, les décès et la migration nette tirées d'une combinaison d'autres sources d'enquête.

3.2 Désaisonnalisation

Un bon nombre d'estimations produites à partir des données de la CPS sont désaisonnalisées selon la méthode X-11-ARMMI qui distingue les changements saisonniers des changements effectivement survenus sur le marché du travail.

4. ÉVALUATION DE LA QUALITÉ DE LA CPS

4.1 Évaluation de l'erreur de réponse

Pour les besoins de la CPS actuelle, les estimations de l'erreur de réponse sont limitées à la variance de réponse. Les données sont recueillies aux fins de cette variance à l'occasion d'une réinterview mensuelle pour détecter l'erreur de réponse. Ainsi, on procède à une sélection aléatoire des cas, et le questionnaire est administré de la même manière que lors de l'interview initiale, sauf que les ménages ayant fait l'objet d'une interview sur place sont habituellement réinterviewés par téléphone. La variance de réponse cumulative pour un an est ensuite analysée afin de déterminer s'il s'est produit des changements par rapport aux estimations passées pour chaque question ou chaque variable composite, par exemple un recodage important de l'emploi. Les données servent aussi à identifier les questions qui ont besoin d'être améliorées à l'occasion de la prochaine refonte du questionnaire.

4.2 Qualité du codage de l'activité économique et de la profession

Au-delà du contrôle de la qualité du travail des codeurs abordé au paragraphe 3.4.1, les données sur la qualité du codage sont également analysées afin d'identifier les codes problèmes. Les résultats servent à mettre à jour le codage assisté par ordinateur dont se servent les codeurs.

4.3 Ratios de couverture

Pour mesurer la couverture, on peut estimer la population d'un groupe avant que les poids soient rajustés au moyen des chiffres de contrôle relatifs à la population, puis diviser cette estimation par le chiffre de contrôle pour ce groupe. Lorsqu'on utilise cette méthode, la CPS tend à couvrir à peu près 93 % de la population civile non institutionnalisée âgée de 16 ans et plus. Par contre, parmi les divers sous-groupes au sein de la population répartis selon l'âge, le sexe, la race et l'ethnicité, la couverture varie de 66 % (mâles noirs âgés de 20 à 29 ans) à 102 % (mâles d'une autre race âgés de 70 ans et plus).

4.4 Taux de non-réponse

Le taux de non-réponse est un indicateur du biais possible lié à la non-réponse. À l'instar de toutes les enquêtes périodiques, la CPS a vu le nombre de refus s'accroître progressivement au fil des ans, bien que ce phénomène ait été neutralisé, dans une certaine mesure, par l'amélioration des méthodes de contact. La CPS affiche actuellement un taux de non-réponse de 7 %.

5. PROJETS D'AVENIR

5.1 Étude d'appariement de la CPS avec le recensement de 2000

Nous procédons actuellement à une analyse de la qualité des données de la CPS à partir de données tirées du recensement de 2000. Cette analyse nous aidera à évaluer l'exactitude de chaque aspect de la qualité de nos données. L'appariement CPS/recensement cherchera à mesurer la couverture des unités de logement et au sein des unités de logement pour les besoins de la CPS, à obtenir des données du recensement pour les non-répondants à la CPS afin d'améliorer notre rajustement pour tenir comte de la non-réponse, et à comparer les réponses données à des questions semblables dans le cadre de la CPS et du recensement.

5.2 Analyse du modèle de structure latente

Parmi les outils très prometteurs d'analyse et de rajustement de la qualité des données, on retrouve l'analyse de structure latente (Biemer et Bushery, 2000). Le modèle markovien de structure latente peut utiliser la conception du panel de la CPS pour estimer l'erreur de réponse, répartie selon les probabilités d'erreur de classification. Nous songeons aussi à utiliser l'analyse de structure latente pour estimer l'effet du biais de renouvellement, répertorié depuis les années 1970 (Bailar, 1975), et possiblement pour rajuster les données afin d'en éliminer l'influence. Bien que la théorie appuie l'efficacité de cette méthode comme estimateur possible de ces facteurs, il nous reste à procéder à des études de validation des concepts, par exemple les simulations et l'appariement avec les dossiers administratifs (notamment avec les donnés du recensement dans l'exemple susmentionné).

6. CONCLUSION

La qualité des données produites à partir de la CPS tient à une combinaison complexe de nombreux facteurs différents. Au-delà du plan de l'enquête axé sur les objectifs de la qualité des données, il est impératif de s'assurer que chaque étape est réalisée avec succès en vue d'atteindre ces objectifs. Pour un projet aussi compliqué que la CPS, les étapes du contrôle de la qualité ne sont pas négligeables et doivent être soigneusement planifiées pour garantir la réussite de l'enquête. Par contre, étant donné chaque niveau de qualité obtenu, le système de contrôle de la qualité devrait servir à l'amélioration continue dans le but de réaliser des objectifs encore plus élevés.

BIBLIOGRAPHIE

Bailar, Barbara A. (1975), "The Effects of Rotation Group Bias on Estimates from Panel Surveys, " *Journal of the American Statistical Association*, vol. 70, p. 23 à 30.

Biemer, Paul P. et Bushery, John M. (2000), "Validité de l'analyse Markorienne de la structure latent pour l'estimation de l'erreur de classiifcation des données de la population active," *Techniques d'enquête*, vol. 26, n° 2, p. 157 à 171.

Fay, Robert et Train, George (1995), "Aspects of survey and model-based postcensal estimation of income and poverty for states and counties," *American Statistical Association Proceedings of the Section on Government Statistics*, p. 154 à 159.

Killion, Ruth A. (1993), "The Impact of Housing Unit Coverage," Housing Unit Coverage Study Results Memorandum Number 2, mémoire du Bureau of the Census de Killion à Walsh daté du 24 juin 1993.

Lyberg, Lars et Kasprzyk, Daniel (1991), "Data Collection Methods and Measurement Error: An Overview," dans P. P. Biemer et coll. (directeurs de publication), *Measurement Errors in Surveys*, New York: Wiley, p. 237 à 257.

U. S. Bureau of the Census, *Technical Paper 63, Current Population Survey, Design and Methodology*, mars 2000