Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Utilisation de dossiers administratifs pour évaluer les données d'enquête

Mary H. Mulry, Elizabeth M. Nichols et Jennifer Hunter Childs[1]

Résumé

Après le Recensement de 2010, le U.S. Census Bureau a mené deux projets de recherche distincts, en vue d'apparier des données d'enquête et des bases de données. Dans le cadre d'une étude, on a procédé à un appariement avec la base de données du tiers Accurint, et dans un autre cas, avec les fichiers du National Change of Address (NCOA) du U.S. Postal Service. Dans ces deux projets, nous avons évalué l'erreur de réponse dans les dates de déménagement déclarées en comparant les dates de déménagement autodéclarées et les enregistrements de la base de données. Nous avons fait face aux mêmes défis dans le cadre des deux projets. Le présent document aborde notre expérience de l'utilisation des « mégadonnées », en tant que source de comparaison pour les données d'enquête, ainsi que les leçons que nous avons apprises pour des projets futurs similaires à ceux que nous avons menés.

Mots-clés : erreur de remémoration; mémoire; télescopage; date du déménagement

1. Introduction

Les dossiers administratifs fournissent une source de données pour évaluer les erreurs dans les réponses aux enquêtes. Ces évaluations peuvent aider à la conception de la collecte des données, mais peuvent aussi fournir une piste pour la conception d'une méthode d'estimation qui dépend d'une combinaison de dossiers d'enquête et de dossiers administratifs, ou une transition des données d'enquête à une source administrative. Même si les dossiers administratifs comprennent des quantités considérables de données, ils sont compilés à leurs propres fins et comportent leurs propres sources d'erreurs. L'utilisation de dossiers administratifs pour évaluer les données d'enquête n'est pas toujours aussi facile qu'on le pense. Dans cet article, nous tirons parti de notre expérience dans le cadre de deux études d'erreurs dans des rapports d'enquête sur les déménagements reposant sur des dossiers administratifs pour aborder certains des défis que doivent relever les chercheurs lorsqu'ils utilisent des dossiers administratifs pour évaluer l'erreur d'enquête (Nichols, Mulry et Childs *à paraître en 2017*). Ces deux études sont un petit élément du grand programme de recherche du U.S. Census Bureau visant à augmenter l'utilisation des dossiers administratifs et des sources de données de tiers dans ses produits d'enquête et de recensement (O'Hara, 2014).

Pour mettre en contexte, les États-Unis ont un système statistique décentralisé. Le U.S. Census Bureau est l'organisme statistique fédéral le plus important, mais d'autres organismes fédéraux, d'état et locaux recueillent et conservent aussi des données. Certains des organismes fédéraux sont statistiques, mais d'autres ont des missions qui ne sont pas statistiques par définition, par exemple, le U.S. Postal Service. De plus, d'autres entités, y compris des entités commerciales ou de tiers et des universités, recueillent et conservent aussi des données. Chacun d'eux recueille des données à ses propres fins, et non pas aux fins du Census Bureau.

Pour utiliser les données d'autres organisations, les responsables du U.S. Census Bureau concluent des ententes avec ces autres entités sur la façon dont ces données seront utilisées et protégées. Il faut souvent des années pour élaborer ces ententes. Il existe des restrictions matérielles et juridiques quant à l'utilisation de ces données, lorsqu'elles sont en possession du U.S. Census Bureau (Johnson, Massey, O'Hara, 2015).

---

[1]Mary H. Mulry, Elizabeth M. Nichols et Jennifer Hunter Childs, U.S. Census Bureau, Washington, DC 20233. Le présent rapport est diffusé pour informer les parties intéressées et pour encourager la discussion concernant les travaux en cours. Les points de vue exprimés sur les questions statistiques, méthodologiques et opérationnelles sont ceux des auteurs et ne reflètent pas forcément ceux du U.S. Census Bureau.

Le U.S. Census Bureau utilise des dossiers administratifs pour la production de certaines de ses estimations économiques et démographiques. Toutefois, les États-Unis ne disposent pas d'une liste continuellement mise à jour de personnes résidant au pays. Le recensement décennal des États-Unis a lieu tous les dix ans et est fondé sur les déclarations des personnes, et non pas sur des dossiers administratifs. Les progrès plus récents dans la technologie informatique et la capacité d'entreposage électronique ont mené à une expansion de la recherche du U.S. Census Bureau quant à l'utilisation de dossiers administratifs et de bases de données de tiers, avec comme objectif de réduire les coûts, d'augmenter l'efficacité, d'améliorer la qualité, de générer de nouveaux produits et de réduire le fardeau de réponse.

2. Thème de la recherche

Nos travaux concernant les deux études sur l'utilisation de dossiers administratifs et de données de tiers à des fins d'évaluation ont commencé par la question de recherche suivante :

> Quelle est l'erreur de mesure dans les rapports d'enquête sur les dates de déménagements résidentiels, au fur et à mesure que le délai entre le déménagement et l'interview augmente?

La réponse à cette question est importante à la fois pour le U.S. Census Bureau et pour les chercheurs d'enquête. Dans le cadre du recensement décennal des États-Unis, on suppose que les personnes qui n'ont pas répondu initialement se rappellent où elles vivaient le jour du recensement (1er avril de l'année de recensement), lorsque les intervieweurs chargés d'assurer le suivi de la non-réponse communiquent avec elles, deux ou trois mois plus tard. En outre, des activités d'évaluation ont lieu plusieurs mois plus tard, selon cette hypothèse. Les chercheurs d'enquête ont formulé des hypothèses similaires concernant la remémoration par les répondants des dates de déménagement, lorsqu'ils utilisent ces événements comme point d'ancrage pour aider à la remémoration d'événements particuliers, plus spécialement dans le contexte qui utilise la technique créant un calendrier de l'historique des événements (Belli, 1998).

Le présent document est axé sur deux projets méthodologiques dans le cadre desquels nous avons tenté de répondre à la question de recherche. Nous avons utilisé des dossiers administratifs et des données de tiers comme source de comparaison avec un rapport d'enquête. Dans un sens, la présente communication est axée sur la faisabilité de ce type d'évaluation et sur ce que nous avons appris au cours de la mise en œuvre de nos études, par opposition à la présentation des résultats auxquels nous sommes arrivés.

Le premier projet a été mené par l'entremise d'un contrat avec NORC (Krishnamurty, 2012). Nous avons ciblé et utilisé trois années de la National Longitudinal Survey of Youth (NLSY) comme source de données d'enquête. La NLSY est une enquête longitudinale annuelle sur place, dans laquelle on demande aux répondants les dates de tous les déménagements interurbains depuis l'interview précédente. Lorsque l'intervieweur a de la difficulté à trouver le répondant échantillonné pour la prochaine vague d'interviews, il a recours à une base de données de tiers appelée Accurint, un outil de localisation et de recherche appartenant à Lexis-Nexus, pour aider à localiser les personnes (Mulry, Nichols, Childs, 2014). Accurint comprend des dossiers pour plus de 400 millions d'identificateurs et compile des données à partir de plus de 10 000 sources. Ces dossiers permettent d'associer le nom d'une personne à une adresse et à des dates de résidence à cette adresse.

Le deuxième projet a été mené au U.S. Census Bureau. Dans le cadre de ce projet, nous avons procédé à un échantillonnage à partir d'une source de dossiers administratifs, puis nous avons mené des interviews auprès des personnes comprises dans l'échantillon. La source était les fichiers du National Change of Address (NCOA) du U.S. Postal Service (USPS; U.S. Postal Service, 2015). Nous avons sélectionné un échantillon de fichiers de formulaires du NCOA en mars et avril 2010, l'année du recensement, et nous avons mené une interview téléphonique auprès des ménages, les mois subséquents, en simulant le moment des différentes opérations de recensement.

Dans le cadre de cette interview, nous avons posé des questions concernant les autres endroits où chaque membre du ménage avait vécu pendant l'année, et les dates de déménagement. Dans ce projet, nous avons aussi comparé les dates du fichier de changements d'adresse aux dates de déménagement déclarées. Nous avons fait face à des défis similaires

dans le cadre des deux projets, et ceux-ci sont pertinents pour d'autres études utilisant des dossiers administratifs comme source de données pour une évaluation des réponses aux enquêtes.

Lorsque nous déterminons si une source de données devrait être utilisée dans une évaluation, l'«adaptation des données à leur utilisation » présente toujours un défi. Dans le cadre de nos deux enquêtes, nous avons posé des questions concernant les dates de déménagement résidentiel, mais une date figurant dans la base de données Accurint ne représentait pas nécessairement une date de déménagement, non plus que la date du NCOA correspondait à un réacheminement du courrier. Toutefois, nous avons dû déterminer si cela faisait une différence pour notre question de recherche. Nous avons conclu que, dans cette étude d'évaluation, les défis liés à la définition sont demeurés constants au fil du temps et, selon cette hypothèse, nous étions toujours sûrs que nos résultats selon lesquels le délai écoulé depuis le déménagement avait des répercussions sur le rappel de ce déménagement étaient toujours valables. Nous ne mesurions pas le nombre de déménagements faits par une personne à partir de ces dossiers, mais si nous l'avions fait, notre conclusion aurait peut-être été différente.

3. Appariement

L'une des premières choses que nous avons dû faire pour les deux projets a été de coupler ou d'apparier les données administratives ou les données de tiers aux données d'enquête. Cela a posé l'un des défis les plus importants des deux études.

Tout d'abord, nous mettons l'accent sur le projet Accurint/NLSY. L'une des principales raisons pour laquelle nous avons sélectionné la base de données Accurint pour notre évaluation est le succès obtenu par NORC dans la gestion de l'attrition, en vue de la maintenir faible, grâce à l'utilisation d'Accurint comme source de localisation pour les répondants de l'échantillon de la NLSY. Notre étude a utilisé l'appariement administratif assisté par ordinateur pour trouver dans la base de données Accurint les dossiers des répondants de la NLSY ayant déménagé qui correspondent aux adresses où ces personnes vivaient depuis la dernière interview.

Initialement, nous avons échantillonné 2 999 répondants à la NLSY, représentant 4 105 déménagements pour les trois cycles de la NLSY (en combinant les déménagements à l'adresse au moment de l'interview et les déménagements dans l'intervalle depuis la dernière interview). La NLSY comportait l'adresse postale pour les déménagements à l'adresse actuelle, étant donné que c'est là que la plupart des interviews sur place ont eu lieu. Toutefois, dans le cadre de la NLSY, on a recueilli uniquement la ville, l'état et le code postal américain pour les déménagements dans l'intervalle depuis l'interview précédente.

Nous avons déterminé que les résultats de l'appariement des répondants de la NLSY et d'Accurint variaient selon le nombre de correspondances entre les dossiers que nous souhaitions obtenir. En appariant le nom seulement, 90 % des répondants ont été retrouvés dans Accurint. Après avoir ajouté l'exigence d'appariement de l'état, le taux d'appariement a diminué pour passer à 69 %. Lorsque l'appariement a été fondé sur le nom, la ville, l'état et le code postal américain à cinq chiffres, le taux d'appariement a diminué à nouveau pour s'établir à 56 %. La baisse du taux d'appariement laisse supposer que le taux d'appariement erroné pourrait être excessivement élevé dans les appariements moins restreints.

Pour les déménagements appariés comportant une correspondance pour le nom, la ville, l'état et le code postal américain à cinq chiffres, la différence entre les dates de déménagement d'Accurint et de la NLSY était beaucoup plus grande que prévue. Ce résultat nous a fait nous interroger sur le fait que les dossiers couplés représentaient le même déménagement. Par conséquent, nous avons décidé d'apparier les personnes ayant déménagé à l'adresse postale où l'interview a eu lieu. Nous avions près de 6 000 adresses d'interview pour nos répondants à la NLSY. Notre taux d'appariement dans le cadre de la deuxième tentative d'appariement a été très similaire à la ronde précédente, soit 57 %, mais nous étions sûrs d'avoir la bonne adresse à ce moment-là. Toutefois, seulement 19 % des adresses de l'interview originale comportaient des dates de déménagement, tant pour l'interview de la NLSY que dans Accurint. Nous croyons que cela peut venir du fait que la base de données Accurint ne couvre pas bien certaines populations, ou que certains déménagements n'ont pas produit un enregistrement pouvant être acquis par Accurint, comme des factures de service public, de cartes de crédit, etc.

Passons maintenant à notre étude NCOA/Recensement, que nous avons entreprise de façon différente de l'étude Accurint/NLSY. Dans cette étude, nous avons commencé par le fichier de dossiers administratifs NCOA, puis nous avons mené les interviews. Nous avons sélectionné 13 500 unités de logement qui ont soumis un changement d'adresse pour mars ou avril 2010 et pour lesquelles nous avons pu trouver des numéros de téléphone dans une base de données de tiers. Le U.S. Census Bureau a mené des interviews téléphoniques permettant de recueillir l'adresse de la résidence, celle de toutes les personnes vivant dans la résidence, et celle d'autres endroits où elles avaient vécu pendant l'année, y compris les dates de déménagement. Dans le cadre de l'étude, on a eu recours à l'appariement administratif entre les dossiers du NCOA et les interviews d'enquête.

Dans le cadre de l'étude NCOA/Recensement, nous avons obtenu un taux de réponse de 66 % globalement, selon le taux de réponse 2 de l'AAPOR (American Association of Public Opinion Research, 2011), qui comprend les unités de l'échantillon dont l'admissibilité est inconnue dans le dénominateur. Toutefois, lorsque nous avons examiné les données plus en détails, nous avons déterminé que seulement 25 % environ de l'échantillon original comportait un nom et correspondait au nom et à l'adresse de réacheminement du fichier NCOA. Par ailleurs, seulement 15 % de l'échantillon original semblait correspondre au même ménage et avait déclaré un déménagement.

Nous croyons que deux facteurs possibles peuvent avoir contribué au faible pourcentage de déménagements déclarés par le même ménage. Peut-être que le numéro de téléphone que nous avions pour le ménage était erroné. Cela peut venir du fait que le couplage pour trouver un numéro de téléphone pour l'adresse de réacheminement a permis de trouver un numéro correspondant à un résident précédent, ou encore que le lien était correct, mais que la personne dont le nom figurait dans le formulaire du NCOA avait déménagé à une autre adresse avant de recevoir un appel dans le cadre de notre étude. L'autre problème est venu du fait que le répondant d'enquête n'avait pas déclaré de déménagement, parce qu'il avait oublié de le faire ou parce que le classement dans le NCOA ne correspondait pas à un déménagement.

4. Localisation du même événement

Les événements que nous avons étudiés étaient des déménagements résidentiels. Même si le déménagement n'est pas un événement qui se produit fréquemment, le U.S. Census Bureau estime qu'environ 12 % de la population des États-Unis déménage chaque année (U.S. Census Bureau, 2011). Dans nos deux études, nous avons trouvé une grande variabilité entre les deux sources en ce qui a trait à la date de déménagement.

Pour l'étude NLSY/Accurint, lorsque nous avons examiné la différence entre la date de début d'Accurint pour l'adresse et la date du déménagement à l'adresse comprise dans le rapport de la NLSY, nous avons noté des écarts considérables. Parfois, la date de début d'Accurint était antérieure à l'interview, et parfois, la date de début se situait après l'interview. Par exemple, certaines personnes dans la NLSY ont indiqué qu'elles avaient déménagé à l'adresse le même mois que l'interview, mais elles étaient couplées à un dossier d'Accurint 50 mois plus tôt. Nous avons conclu que nous étions en présence du mauvais événement. Autrement, cela laisserait supposer qu'Accurint est au courant du déménagement d'une personne quatre ans à l'avance. Un exemple extrême s'est produit pour deux répondants qui ont indiqué avoir déménagé à l'adresse quatre mois plus tôt, mais qui ont été couplés au dossier Accurint environ 25 ans plus tôt. Il faut se rappeler que les répondants à la NLSY étaient âgés entre 23 et 29 ans, ce qui fait que l'hypothèse la plus probable est que leurs dossiers Accurint ont été associés à leur naissance. À notre connaissance, Accurint n'effectue pas de contrôle ou de correction des dates, ce qui entraîne les écarts que nous avons observés.

Certains des liens comportant des différences importantes correspondaient à l'adresse des parents, mais ce n'était pas le cas pour la majorité. Afin de tenter d'obtenir l'événement approprié, nous avons dû établir des sous-ensembles de nos données et nous avons examiné uniquement les appariements se situant dans un délai particulier l'un de l'autre (p. ex, un à deux ans), puis nous avons examiné comment les résultats ont changé lorsque cette période a été modifiée.

Dans l'étude NCOA/Recensement, il a été plus facile d'obtenir l'événement approprié. Dans l'enquête, nous avons recueilli des dates et nous avions presque 2 000 dates de déménagement déclarées. Environ 200 d'entre elles concernaient des situations où la personne avait déclaré déménager à l'adresse de réacheminement, mais n'avait pas déclaré le mois du déménagement à cette adresse. En l'absence de la date de déménagement, ces cas ont été laissés de côté dans notre analyse.

5. Généralisation des résultats

Comme vous pouvez l'imaginer, diviser les données en sous-ensembles a eu des répercussions sur la façon de généraliser nos résultats. Toutefois, nos analyses dans les deux études, bien qu'elles ait été limitées, ont permis de déterminer une augmentation plus grande des erreurs de remémoration des dates de déménagement, à partir de six à dix mois après le déménagement.

Pour l'étude NLSY/Accurint, nos répondants avaient de 23 à 29 ans, en raison de la cohorte visée par la NLSY. Même si nous avons commencé avec près de 3 000 répondants, à la fin, nous n'avons pu utiliser que 410 déménagements pour l'analyse. Il s'agissait d'événements pour lesquels nous pouvions confirmer avoir la bonne personne, la bonne adresse et le bon déménagement. Nous n'avons pas été en mesure de trouver des contrôles appropriés pour la pondération des personnes ayant déménagé dans ce groupe d'âge, ce qui a rendu muette la question de la généralisation.

L'étude NCOA/Recensement a commencé par un ensemble de données comprenant les personnes ayant déménagé, qui avaient enregistré un changement d'adresse. Toutefois, le NCOA ne comprend pas toutes les personnes ayant déménagé (parce que ce ne sont pas toutes ces personnes qui produisent un changement d'adresse auprès du USPS) et que toutes les personnes qui produisent un changement d'adresse ne déménagent pas nécessairement (on peut concevoir des raisons de faire réacheminer son courrier, lorsque l'on ne déménage pas dans les faits). Nous avons commencé avec 13 500 ménages échantillonnés et terminé avec 1 740 déménagements. Encore une fois, nous n'avions pas de contrôles appropriés pour pondérer les personnes ayant déménagé en mars et avril 2010. Les données historiques que nous avons pu trouver ne semblaient pas appropriées pour la pondération de nos données.

6. Fréquence des mises à jour des dossiers

Le moment de la mise à jour des données administratives ou de tiers a eu des répercussions sur notre analyse.

Le flux de l'étude NLSY/Accurint découlait du fait que les interviews de la NLSY que nous avons utilisées se sont déroulées d'octobre 2006 jusqu'à la fin de mai 2009. Le NORC a tiré les données d'Accurint en 2011 pour ces cas, puis l'appariement a été effectué. Nous n'étions pas certains du moment de notre évaluation, mais a posteriori, cela a très bien fonctionné. La recherche dans Accurint aussi longtemps après les interviews a donné suffisamment de temps pour que celui-ci soit mis à jour et a permis de trouver un appariement pour certain des événements qui ont été entrés dans Accurint des mois, voire des années, après l'interview. Si nous avions effectué l'évaluation plus près de l'interview, la base de données de tiers n'aurait pas comporté autant de mises à jour et de dossiers ajoutés, ce qui signifie qu'un nombre encore moins grand de cas aurait été disponible en vue de l'inclusion dans notre analyse.

Dans l'étude NCOA/Recensement, la période de référence des dossiers administratifs utilisés a peut-être eu un effet sur les taux de réponse. Les dossiers du NCOA pour mars et avril 2010 ont été tirés, puis les adresses de réacheminement ont été appariées aux numéros de téléphone, à partir d'une base de données administratives de tiers différente. À partir de ces numéros de téléphone, nous avons mené les interviews en juin, septembre et février. Nous avons obtenu un taux de réponse de 69 % en juin, qui est passé à 63 % en février. Nous croyons que la baisse est attribuable au fait que les adresses ont été appariées à des numéros de téléphone en mai, et que certains de ces numéros étaient périmés au mois de février suivant.

Lorsque l'on effectue ces types d'études, on doit garder en tête l'erreur dans les bases de données. Les sources possibles d'erreur dans les bases de données comprennent les suivantes :

- les données déclarées par une personne ou un ménage peuvent être de meilleure qualité que celles provenant d'un magazine ou d'un service public (la source de certains dossiers Accurint);
- les différences de définition, comme l'adresse temporaire par rapport à l'adresse permanente;
- une résidence au sens de la loi par rapport à la résidence habituelle (où la personne vit habituellement); et
- la qualité de certains renseignements, qui peut être plus importante pour le propriétaire de la base de données que d'autres types. Par exemple, la date de naissance est très importante pour la U.S. Social Security Administration, étant donné que l'âge de la personne détermine l'admissibilité aux prestations.

Par conséquent, un chercheur doit connaître les répercussions des sources de données, des modèles de mise à jour et des priorités des données lorsqu'il choisit une base de données pour la recherche sur une enquête déclarant des erreurs dans la variable d'intérêt. Une base de données peut représenter une bonne source pour certains types de variables, certains mois, mais pas toute l'année.

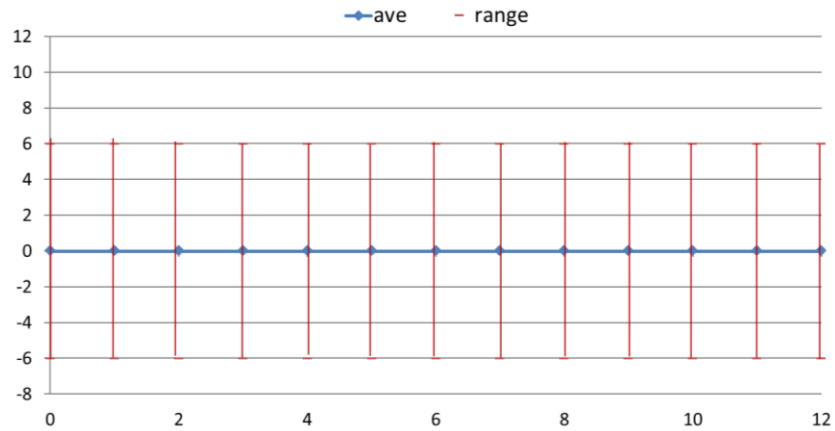
Dans les études où la base de données sert de base de sondage pour l'échantillonnage des personnes qui ont déménagé, la mise à jour de la base de données a des répercussions sur l'étude, à l'étape de la sélection de l'échantillon, qui prend la forme d'un sous-dénombrement des personnes qui ont déménagé, comme dans l'étude Recensement/NCOA. Dans le cas du fichier du NCOA, il peut y avoir surdénombrement parce que certains changements d'adresse ne correspondent pas à des déménagements.

Lorsque la base de données n'est pas la base d'échantillonnage, mais plutôt la source de comparaison pour le rapport d'enquête sur les déménagements, comme c'était le cas dans l'étude NLSY/Accurint, les problèmes entourant la mise à jour de la base de données ont des répercussions sur la présence de l'événement dans cette base. Nous illustrons ci-après deux modèles d'erreurs dans les bases de données, à partir d'erreurs dans les adresses des personnes ayant déménagé, à savoir le moment où la base de données a acquis la nouvelle adresse associée à un déménagement. Les adresses pour les personnes ayant déménagé représentent probablement l'un des types de données les plus sujets aux erreurs.

La figure 6.1 comprend une illustration d'une base de données qui fait continuellement l'acquisition de données et les met à jour sur une base mensuelle. Cette illustration repose sur le principe qu'une nouvelle adresse est utilisée au plus six mois avant le déménagement. Pour cette base de données, la durée moyenne d'acquisition de la nouvelle adresse est de zéro, ou le mois du déménagement. Toutefois, il existe un large éventail du nombre de mois qu'il faut pour que la base de données acquière l'adresse.

Nous croyons qu'Accurint affiche un modèle similaire à celui compris dans la figure 6.1. Nous n'avons pas décelé de biais dans la durée d'acquisition d'une nouvelle adresse par Accurint, mais nous disposions d'une petite quantité de données et nous avons fait des hypothèses dans nos analyses. La durée d'acquisition stable au fil du temps peut faire en sorte qu'une base de données est une candidate raisonnable pour l'évaluation de l'erreur de déclaration, lorsque des données sont recueillies sur une période, par exemple la période de trois ans de notre étude NLSY/Accurint. Il existe une autre considération, à savoir si la stabilité dans le délai d'acquisition global de la base de données existe aussi pour la sous-population d'intérêt et les données utilisées dans l'analyse.

Figure 6.1.
Illustration : Délai nécessaire pour que la nouvelle adresse d'une personne qui déménage apparaisse dans une base de données qui fait l'acquisition de dossiers auprès de nombreuses sources chaque mois.

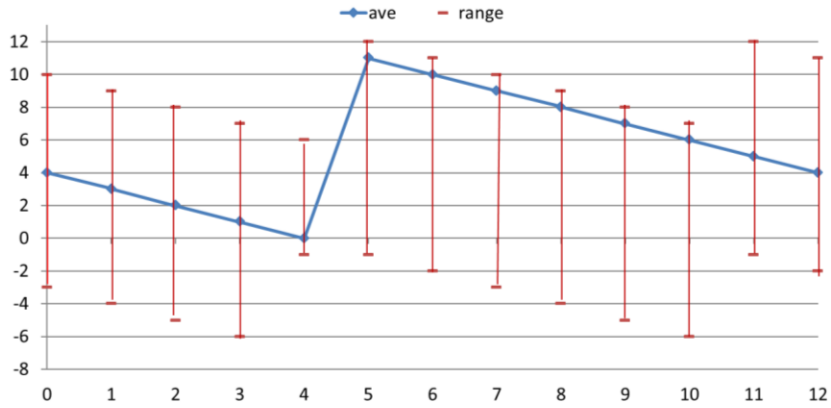


Hypothèses : La base de données fait continuellement l'acquisition de nouveaux dossiers.
Une nouvelle adresse est utilisée au plus six mois avant le déménagement et est mise à jour au plus tard six mois après le déménagement.

Par contre, la figure 6.2 comprend la deuxième illustration d'un autre type de base de données, qui procède à une mise à jour une fois par année, au mois 4, mais recueille aussi quelques mises à jour au mois 10. L'idée principale est que la base de données est mise à jour chaque année, et non pas mensuellement. Encore une fois, cette figure rend compte d'une conjecture concernant les tendances d'acquisition d'une nouvelle adresse par une telle base de données. La figure 6.2 montre que le mois 4, le mois de la mise à jour, comporte le délai moyen le plus court pour qu'une adresse se retrouve dans la base de données. Le délai nécessaire pour l'acquisition de l'adresse n'est pas symétrique concernant la moyenne, tandis que l'erreur dans la base de données de la figure 6.1 est symétrique.

La base de données de la figure 6.2, comportant des mises à jour une fois par année, montre une tendance d'erreur pour les adresses des personnes qui déménagent très différente de la base de données précédente comprenant une acquisition continue de données et des mises à jour mensuelles. La caractéristique importante de cette base de données est que le délai d'acquisition d'une nouvelle adresse varie pendant l'année, ce qui fait qu'elle n'est pas une bonne candidate pour l'évaluation des rapports d'enquête sur les mois de déménagement. Toutefois, elle pourrait être une bonne candidate pour l'évaluation d'autres variables, comme l'adresse d'une personne le mois 4 ou les variables annuelles, comme le revenu annuel.

Figure 6.2.
Illustration : Délai nécessaire pour que la nouvelle adresse d'une personne qui déménage apparaisse dans une base de données qui fait l'acquisition de presque tous ses dossiers au quatrième mois



Hypothèses : Quelques dossiers sont acquis au mois 10, mais très peu à d'autres moments.
Une nouvelle adresse est utilisée au plus six mois avant le déménagement.

7. Résumé

Lorsque l'on planifie une étude, le biais et l'erreur aléatoire dans la base de données pour la variable d'intérêt représentent une considération importante. Les propriétés d'erreur de la base de données peuvent avoir des répercussions :

- sur le moment des interviews de l'enquête, c'est-à-dire la tenue des interviews le plus près possible des mises à jour de la base de données;
- la méthodologie choisie pour l'analyse.

En résumé, nous avons appris certaines choses de nos deux expériences, qui s'appliquent à d'autres études reposant sur des dossiers administratifs ou des bases de données de tiers pour évaluer l'erreur dans les rapports d'enquête :

- planifier l'appariement soigneusement, particulièrement pour les événements qui se produisent fréquemment;
- connaître les définitions et les limites de toutes les données;
- planifier le moment où les données sont tirées de façon optimale;
- être prêt à obtenir un ensemble de données beaucoup plus petit que prévu;
- connaître comment le biais et l'erreur aléatoire de la base de données choisie ont des répercussions sur votre population d'intérêt, votre plan d'étude et la capacité de généraliser vos résultats.

Bibliographie

American Association for Public Opinion Research (2011), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition*. AAPOR. http://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/StandardDefinitions2011_1.pdf (Version du 6 février 2015).

Belli, R. F. (1998), The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory, 6*, 383-406.

Johnson, D. S., C. Massey, et A. O'Hara (2015), "The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility," *The ANNALS of the American Academy of Political and Social Science January 2015*, 657. pp. 247-264.

Krishnamurty, P. (2012), "Memory Recall of Migration Dates in the National Longitudinal Survey of Youth, 1997 Cohort: Statistical Analysis and Evaluation" (Census R&D 2014 IDIQ Contract No. YA 1323-09-CQ-0053). NORC.

Linse, K., T. Pape, L. Rosenberger, et G. Contreras (2010), Census Coverage Measurement Survey Recall Bias Study. U.S. Department of Commerce, U.S. Census Bureau, Washington, DC.

Mulry, M. H. (2014), Measuring Undercounts for Hard-to-Survey Groups. In R. Tourangeau, N. Bates, B. Edwards, T. Johnson, et K. Wolter (Eds.), *Hard-to-Survey Populations.* (Chapter 3). Cambridge University Press, Cambridge, England. pp. 37 – 57.

Mulry, M. H., E. M. Nichols, et J. H. Childs (2014), "Study of error in survey reports of move month using the U.S. Postal Service Change of Address records. In *JSM Proceedings*, Survey Research Methods Section. American Statistical Association, Alexandria.

Mulry, M. H., E. M. Nichols, et J. H. Childs (2015), "Evaluating recall error in survey reports of move dates through a comparison with records in a commercial database." Article non publié. U.S. Census Bureau. Washington, DC.

Nichols, E. M., M. H. Mulry, et J. H. Childs (*Forthcoming in 2017*), "Using administrative records data at the U.S. Census Bureau: Lessons learned from two research projects evaluating survey data." In Biemer, P.P, Eckman, S., Edwards, B., Lyberg, L., Tucker, C., de Leeuw, E., Kreuter, F., et West, B.T. (Eds.), *Total Survey Error in Practice*. Wiley. New York.

O'Hara, A. (2014), "Comments on: Laying the foundations for a new approach to Census taking in Ireland by John Dunne and Steve MacFeely." *Statistical Journal of the IAOS, 30.* pp. 367-368.

U.S. Census Bureau (2011), *Mover Rate Reaches Record Low, Census Bureau Reports*. Dernière version accédée: 28 août 2012: http://www.census.gov/newsroom/releases/archives/mobility_of_the_population/cb11-193.html

U.S. Postal Service, NCOA^{Link} (2015), [WWW document]. Dernière version accédée: 7 août 2015: https://ribbs.usps.gov/index.cfm?page=ncoalink