Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

INFÉRENCES AYANT TRAIT À DES POPULATIONS FINIES D'APRÈS DES DONNÉES PROVENANT DE SOURCES MULTIPLES ET DONT LES PÉRIODES DE RÉFÉRENCE DIFFÈRENT

Jana Asher[1], Stephen E. Fienberg[2], Elizabeth Stuart[3] et Alan M. Zaslavsky[4]

RÉSUMÉ

On considère souvent que les recensements et les enquêtes donnent des mesures des populations telles qu'elles sont, alors qu'en fait, la plupart reflètent les renseignements sur les individus tels qu'ils étaient au moment où a été faite la mesure, voire même à un point antérieur dans le temps. Par conséquent, les inférences faites à partir de telles données doivent tenir compte des changements qui surviennent au cours du temps au niveau tant de la population que de l'individu. Plus précisément, au niveau de la population, les variables étudiées, telles que la taille ou les caractéristiques moyennes d'une population, pourraient évoluer. Parallèlement, des sujets individuels pourraient rentrer dans le champ de l'étude ou en sortir, ou changer de caractéristiques. Ces changements au fil du temps peuvent avoir des répercussions sur les études statistiques de données gouvernementales qui regroupent des renseignements provenant de sources multiples, y compris des recensements, des enquêtes et des dossiers administratifs, pratique qui devient de plus en plus courante. Les inférences d'après les bases de données fusionnées résultantes dépendent souvent fortement de choix particuliers faits au moment de combiner, de vérifier et d'analyser les données qui reflètent des hypothèses quant à l'évolution ou à la stabilité de la population au cours du temps. Nous décrivons un premier effort en vue de définir un cadre unique pour ce genre de problème d'inférence, en donnant pour l'illustrer divers exemples dont 1) la combinaison de dossiers administratifs pour estimer la taille de la population des États-Unis, 2) l'estimation de la situation de résidence le jour du recensement d'après des dossiers administratifs multiples, 3) l'estimation de la prévalence de l'abus des droits de l'homme et 4) l'utilisation des moyennes mobiles pour l'American Community Survey.

MOTS CLÉS : dossiers administratifs; American Community Survey; modèles hiérarchiques; méthodes d'estimation à systèmes multiples; modèles stochastiques.

1. INTRODUCTION

Dans la littérature statistique sur les enquêtes par sondage et leur conception, nous soutenons habituellement que les recensements et les enquêtes donnent des mesures des populations telles qu'elles sont, que ce soit en matière de faits ou d'opinions (voir, par exemple, Cochran, 1977; Lyberg et Cassel, 2001). Conjuguée à cette perspective, la théorie classique de l'échantillonnage de population finie traite les mesures comme des valeurs fixes (voir Särndal, Swensson et Wretman, 1992). Pourtant, la réalité diffère habituellement de ces déclarations et de la théorie classique. La plupart du temps, nous recueillons des données sur des individus tels qu'ils étaient à un point antérieur dans le temps, à l'aide de mesures faites encore à un autre moment. Dans un tel contexte, les variations temporelles influent sur les inférences. Au niveau de la population, les variables d'intérêt, comme la taille ou les caractéristiques moyennes d'une population peuvent évoluer. Au niveau de l'individu, des sujets peuvent rentrer dans le champ de l'étude ou en sortir, ou changer de caractéristiques.

Ces changements au fil du temps peuvent avoir des répercussions sur les études statistiques de données gouvernementales qui regroupent des renseignements provenant de sources multiples, y compris des recensements,

[1]Department of Statistics, Carnegie Mellon University, Pittsburgh PA, U.S.A., 15213-3890 (asher@stat.cmu.edu)
[2]Department of Statistics and Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh PA, U.S.A., 15213-3890 (fienberg@stat.cmu.edu)
[3]Department of Statistics, Harvard University, Cambridge MA, U.S.A., 02138 (stuart@stat.harvard.edu)
[4]Harvard Medical School, Department of Health Care Policy, Boston MA, U.S.A., 02115 (zaslavsk@hcp.med.harvard.edu)

des enquêtes et des dossiers administratifs, pratique qui devient de plus en plus courante. Les inférences d'après les bases de données fusionnées résultantes dépendent souvent fortement de choix particuliers faits au moment de combiner, de vérifier et d'analyser les données qui reflètent des hypothèses, habituellement implicites, quant à l'évolution ou à la stabilité de la population étudiée au cours du temps. Pour ancrer solidement ces inférences, nous avons besoin de modèles stochastiques explicites qui saisissent les changements au cours du temps. Ce besoin est reconnu depuis longtemps dans le contexte des enquêtes par panel où l'évolution au cours du temps est le point d'intérêt (p. ex., voir Fienberg, 1980; Stasny, 1990; Pfeffermann et coll., 1998; Ybarra et Lohr, 2002), mais il a été en grande partie ignoré pour des situations où l'on vise à estimer une seule caractéristique d'une population à un point particulier dans le temps ou une série de caractéristiques.

Nous décrivons ici un premier effort en vue de définir un cadre unique pour ce genre de problème d'inférence portant sur des mesures et des systèmes multiples, en donnant pour l'illustrer divers exemples dont :

1. la combinaison de dossiers administratifs pour estimer la taille de la population des États-Unis;
2. l'estimation de la situation de résidence le jour du recensement d'après des dossiers administratifs multiples;
3. l'estimation du nombre d'homicides commis au Kosovo;
4. l'utilisation de moyennes mobiles pour l'American Community Survey.

Dans les sections qui suivent, nous décrivons chacun de ces exemples, puis nous revenons, à la dernière section, à leurs points communs et aux questions plus générales de modélisation.

2. ESTIMATION À SYSTÈMES MULTIPLES DU RECENSEMENT DE 2000

Depuis au moins six recensements décennaux, le U.S. Census Bureau utilise diverses méthodes pour évaluer l'exactitude et la couverture des dénombrements décennaux. En 2000, cette évaluation a pris la forme de l'Accuracy and Coverage Evaluation Survey (ACE), enquête qui a été réalisée auprès d'environ 314 000 unités de logement. Dans le cadre d'un effort distinct, le Bureau a étudié l'utilisation de dossiers administratifs pour remplacer ou pour compléter le dénombrement traditionnel et d'autres enquêtes. Plus précisément, la Planning, Research and Evaluation Division (PRED) du U.S. Census Bureau a mis en place un programme en vue de produire, annuellement, une « superliste » de dossiers administratifs appelée Statistical Administrative Records System (StARS), qui est le produit final de la fusion minutieuse d'au moins six sources de dossiers administratifs. Dans le cadre du programme d'évaluation du Census Bureau pour le Recensement de 2000, la PRED a élaboré une expérience sur des dossiers administratifs (AREX 2000) en vue d'étudier la faisabilité de l'utilisation de StARS pour un recensement sur dossiers administratifs dans cinq comtés du Colorado et du Maryland.

Dans le cadre de l'AREX 2000, on s'est efforcé de mettre au point des méthodes d'estimation à systèmes multiples permettant de combiner les données du recensement décennal de 2000, les données de l'enquête ACE et le fichier ascendant de dossiers administratifs du projet AREX 2000 afin de créer des dénombrements de population au niveau de l'îlot. Plusieurs questions se sont posées durant l'étude. L'une d'elles était la couverture permise par l'ACE; l'appariement sur les trois fichiers crée des classifications croisées tridimensionnelles pour les îlots échantillonnés de l'ACE et des classifications croisées bidimensionnelles pour les îlots non échantillonnés. Il fallait donc concevoir une méthode de modélisation d'ensemble pour ces divers types de données. Un problème plus grave avait trait à la période de référence de chaque liste; les données de l'AREX ont été recueillies bien avant 2000, celles du Recensement de 2000, aux alentours d'avril 2000 et celles de l'ACE, plus tard en 2000. Par conséquent, le regroupement de ces données dans un tableau de contingence s'avérait problématique.

2.1 Couplage d'enregistrements et estimation

La première étape du projet d'estimation à systèmes multiples a été de créer un tableau de contingence des dénombrements de population pour chaque îlot en se servant des trois sources de données décrites plus haut. Comme ces sources de données ont chacune une période de référence différente, il se pourrait qu'un individu ait des adresses différentes sur des listes distinctes et, par conséquent, qu'il soit dans des blocs différents sur des listes distinctes. Pour résoudre ce problème, nous avons supposé que les adresses multiples résultent systématiquement de périodes de référence différentes pour les listes (et non de résidence multiples). Comme le point de référence souhaité était

celui du recensement décennal (1er avril 2000), nous avons placé tous les appariements triples dans leur adresse du Recensement de 2000. Nous avons utilisé une stratégie semblable pour les appariements doubles; les appariements Recensement de 2000-ACE et Recensement 2000-AREX 2000 ont été placés dans leur adresse du Recensement de 2000, mais les appariements ACE-AREX 2000 ont été placés dans leur adresse de l'ACE.

Par la suite, nous avons élaboré plusieurs cadres d'estimation à systèmes multiples pour analyser les tableaux de contingence résultants. Ces cadres intègrent des hypothèses au sujet de l'absence d'îlots échantillonnés dans l'ACE en incluant des stratifications ou des covariables d'après la base d'échantillonnage de l'ACE. La modélisation dans chaque strate de l'ACE permet de ne pas tenir compte des cas manquants; l'inclusion de covariables d'après les caractéristiques selon lesquelles les îlots ont été stratifiées pour l'échantillon de l'ACE permet de tenir compte des cas manquants dans le modèle. Les modèles incluent un ensemble commun de paramètres (dans la strate pour le modèle stratifié) sur tous les îlots pour l'effet de l'ACE et les effets de l'interaction entre l'ACE et les autres listes, ainsi que des paramètres individuels pour chaque îlot pour les effets du Recensement de 2000 et de l'AREX-2000 et pour leur interaction. Les résultats de cet effort de modélisation seront publiés ultérieurement; le lecteur trouvera d'autres renseignements sur les modèles dans Asher et Fienberg (2002).

2.2 Avantages et problèmes

L'un des avantages de ce cadre de modélisation est sa souplesse; ces modèles peuvent être adaptés de façon à refléter non seulement les omissions, mais aussi les inclusions erronées ou les erreurs dans chaque source de données. Ces erreurs sont intrinsèquement stochastiques; au fil du temps, les chiffres de population varient pour des îlots particuliers, à mesure que surviennent des naissances, des décès et des migrations. Comme notre objectif est d'estimer les chiffres de population d'îlot le jour du recensement, nous avons choisi d'extrapoler les données pour les individus de l'AREX et de rétropoler celles pour les individus de l'ACE durant la fusion des données. Autrement dit, nous avons tenu compte de la nature stochastique de ces données lors de la création du tableau de contingence, mais non dans le modèle proprement dit. Dans de futurs travaux, nous essayerons d'intégrer la nature stochastique des données directement dans le modèle.

À cette fin, un problème que pose notre cadre de modélisation tient au fait que les modèles ne sont pas vraiment conçus pour « estimer » des données sur les individus, voire même les ménages, mais pour des agrégats. Par conséquent, l'information stochastique disponible au niveau de l'individu (p. ex., les changements d'adresse) doit être intégrée dans un modèle qui s'appuie sur les classifications croisées au niveau de l'îlot. Une solution éventuelle consiste à produire une classification croisée supplémentaire selon la situation de migration.

3. MODÉLISATION DE LA RÉSIDENCE ET DE LA MIGRATION

La présente section décrit les travaux en vue de modéliser les changements d'adresse au niveau individuel évoqués plus haut. Comme nous l'avons mentionné dans la section 2, ces dernières années, il a été question d'utiliser des dossiers administratifs pour compléter les opérations du U.S. Census Bureau, particulièrement le recensement décennal. L'un des inconvénients des dossiers administratifs est que leur période de couverture ne coïncide pas avec le jour du recensement et peut s'étendre sur une période considérablement antérieure. Étant donné la migration, certains individus qui ne sont plus des résidents le jour du recensement pourraient être inclus dans les enregistrements. Par exemple, quelqu'un pourrait renouveler un permis de conduire à une adresse particulière, mais ne plus donner lieu à aucune autre entrée dans un système d'enregistrement des permis de conduite jusqu'au renouvellement suivant qui doit avoir lieu quelques années plus tard. Les chiffres estimés de recensement qui incluent tous les individus de ce genre seront biaisés. Nous développons un modèle de migration et d'observations dans les dossiers administratifs qui utilisent toute l'information contenue dans un ensemble d'enregistrements, notamment les dates des enregistrements, pour prédire si une personne réside encore à l'adresse indiquée le jour du recensement. De cette façon, nous pouvons estimer les chiffres de population le jour du recensement pour les domaines d'intérêt, en tenant compte du décalage temporel.

3.1 Modélisation hiérarchique stochastique avec migration

Nous présentons ici une vue d'ensemble du modèle; des détails supplémentaires, les algorithmes pour l'ajustement des modèles et les résultats des simulations figurent dans Stuart et Zaslavsky (2001, 2002). Le modèle est une extension de l'estimation à systèmes multiples, qui consiste habituellement à traiter plusieurs systèmes de données comme étant effectivement contemporains, et des méthodes de capture-recapture utilisées pour estimer les populations animales, qui s'appuient sur la « capture » d'individus par des méthodes semblables en plusieurs points dans le temps et peuvent inclure une composante pour la migration des individus dans la zone de capture et hors de celle-ci. Nous supposons que les données comprennent plusieurs systèmes, qui incluent chacun un ensemble d'enregistrements qui reflète la présence d'un individu à une adresse à une date particulière. Les systèmes d'enregistrement peuvent être des systèmes administratifs, un recensement ou une enquête d'évaluation de la couverture. Nous considérons chaque système d'enregistrement administratif ou chaque recensement comme étant une « capture ». Le processus qui génère les enregistrements est représenté par un modèle hiérarchique à trois niveaux que l'on peut décrire de façon générique comme suit. Ce cadre général permet d'adapter une gamme de modèles spécifiques.

Niveau I (données) : Observation dans un système d'enregistrement particulier dans le domaine d'intérêt et dans la période de référence.
Niveau II : Présence de l'individu dans un système particulier à n'importe quel endroit ou moment, et moment où l'individu serait observé s'il est dans le système.
Niveau III : Antécédents de migration de l'individu.

Le niveau III décrit la distribution des dates de migration dans le domaine d'intérêt et hors de celui-ci. Un modèle simple et raisonnablement flexible que nous avons utilisé dans les simulations décrit la population comme étant un mélange de deux types d'individus : les personnes qui ne déménagent pas et celles qui déménagent. Les personnes qui ne déménagent pas ne viennent jamais s'installer dans le domaine d'intérêt ni n'en déménagent jamais, tandis que les personnes qui déménagent sont caractérisées par un hasard constant de partir et d'être remplacées à un taux égal par des personnes qui déménagent dans le domaine d'intérêt. L'hypothèse de hasard constant est probablement une approximation raisonnable sur des intervalles courts, mais on peut aussi utiliser d'autres modèles de migration.

Les niveaux I et II représentent des modèles d'observation spécifiques pour chaque type d'enregistrement. Les variables de niveau II pour chaque personne selon chaque système d'enregistrement incluent une variable indicatrice de l'inclusion de la personne dans le système d'enregistrement et une variable pour la date à laquelle la personne serait enregistrée, si elle était incluse. Donc, les paramètres incluent la probabilité qu'un individu possède ce type d'enregistrement (probabilité qui peut dépendre des valeurs des covariables au moyen d'un modèle de régression) et les paramètres de la distribution des dates associées à ce type d'enregistrement. Par exemple, si le renouvellement des permis de conduire se fait à la date d'anniversaire, les dates de renouvellement figurant dans les enregistrements sont distribuées uniformément sur l'année; en revanche, les dates de production de la déclaration de revenus sont caractérisées par une distribution non uniforme présentant un sommet à l'échéance annuelle de production des déclarations. Le modèle de niveau I décrit si une personne particulière figure dans un système donné en tant que fonction des variables de niveau II et des antécédents de migration (variable de niveau III); cette partie du modèle pourrait être déterministe. Les modèles spécifiques dépendent des détails qui figurent dans les systèmes d'enregistrement. Des exemples figurent dans Stuart et Zaslavsky (2001, 2002). Il est facile de modéliser concurremment les systèmes à enregistrements multiples en les modélisant individuellement aux niveaux I et II, puis en les combinant en supposant que les systèmes sont indépendants conditionnellement aux covariables et aux dates de migration.

Étant donné la structure hiérarchique, le modèle permet plusieurs niveaux d'inférence. Les paramètres globaux, comme les probabilités de couverture et les taux de migration, peuvent être estimés à condition de disposer de données adéquates. Aux fins des dénombrements du recensement, on peut calculer la probabilité a posteriori que chaque individu était un résident le jour du recensement. On peut alors estimer la population au recensement au niveau d'un bloc ou d'un secteur de recensement comme étant le nombre prévu de résidents le jour du recensement.

L'application de ce modèle en se servant à la fois de données simulées et de données provenant d'enregistrements administratifs du Census Bureau a donné des résultats encourageants. Consulter Stuart et Zaslavsky (2001, 2002) pour un résumé des résultats en simulation. Lors de travaux inédits, un ensemble de systèmes d'enregistrement provenant du Statistical Administrative Records System (StARS) du U.S. Bureau of the Census a donné des estimations de population proches des chiffres du recensement décennal. Au niveau individuel, les résultats sont semblables à ceux attendus. Par exemple, il est assez improbable qu'un individu qui n'est observé dans le domaine que dans un fichier d'inscription au régime d'assurance-maladie deux ans avant le jour du recensement soit un résident du domaine le jour du recensement. Par contre, il est fort probable qu'un individu observé dans un fichier d'inscription au régime d'assurance-maladie deux ans avant le jour du recensement et dans un fichier de déclarations de revenus de l' Internal Revenue Service à la même adresse six semaines avant le jour du recensement soit un résident de ce domaine le jour du recensement.

3.2 Avantages et problèmes

Un avantage important de l'approche décrite ici est la modélisation au niveau individuel. Elle permet d'utiliser l'information complète qui figure dans les enregistrements, y compris celle sur les covariables et les dates, et toute l'information sur les adresses pour chaque individu. L'utilisation antérieure des dossiers administratifs par le U.S. Census Bureau reposait sur la sélection de la « meilleure » adresse pour chaque individu, en se fondant sur une série de règles de sélection plus ou moins ponctuelles. Le présent modèle utilise toutes les informations disponibles sur les adresses pour déterminer la probabilité de résidence le jour du recensement pour chaque individu.

Un deuxième avantage important de cette approche est la flexibilité des modèles pour divers systèmes d'enregistrement, ce qui permet de refléter pleinement les caractéristiques de chaque système. Le cadre général permet d'utiliser divers types d'enregistrement, comme les dossiers fiscaux, les fichiers d'inscription au régime d'assurance-maladie et les listes d'enregistrements à des services particuliers. Par exemple, pour modéliser les 1 040 déclarations de revenus de l'Internal Revenue Service, on peut développer un modèle qui décrit la probabilité qu'un individu produise une déclaration de revenus, modèle qui peut dépendre des covariables pour lesquelles des données sont disponibles, comme l'âge, la région du pays, le sexe et la taille de la famille. La distribution des dates de production de la déclaration pourrait être modélisée paramétriquement ou estimée non paramétriquement en se servant de la distribution empirique des dates de production des déclarations. Le modèle de migration est également très général et permet de tenir compte de la migration saisonnière ou des variations régionales des profils de migration. Tous ces modèles peuvent être rendus plus réalistes ou plus complexes, au besoin.

Grâce à l'utilisation d'enregistrements individuels et du moment où ces enregistrements ont été produits pour obtenir des estimations de la taille de la population le jour du recensement, le cadre décrit ici nous offre une approche permettant de tempérer les effets des discordances entre les périodes de référence des systèmes d'enregistrement.

4. HOMICIDES AU KOSOVO, 1999

Dans le cadre de l'accusation portée contre Slobodan Milosevic durant son procès devant le Tribunal pénal international pour l'ancienne Yougoslavie, une équipe de scientifiques américains ont entrepris d'analyser les données disponibles sur les migrations et les décès des personnes d'origine albanienne survenus au Kosovo de mars à juin 1999 (voir Ball et coll., 2002). Le but de cette analyse était de corroborer ou de réfuter trois théories quant à la cause de ces migrations et de ces décès : 1) les actions de l'Armée de libération du Kosovo (ALK) ont causé les décès et les migrations des résidents d'origine albanienne, 2) les attaques aériennes par les forces de l'Organisation du Traité de l'Atlantique Nord (OTAN) ont causé les décès et les migrations des résidents d'origine albanienne et 3) une campagne systématique menée par les forces yougoslaves ont causé les décès et les migrations des résidents d'origine albanienne. Étant donné la nature des questions soulevées par l'accusation, les dates représentaient un aspect essentiel de l'analyse entreprise. Malheureusement, l'information sur ces dates n'étaient pas exacte dans de nombreux cas. L'information sur les migrations provenait de deux sources, à savoir les registres tenus par les gardes à la frontière entre le Kosovo et l'Albanie et les rapports de l'ONU. Les données sur les décès des résidents d'origine albanienne provenaient de quatre sources, établies chacune d'après plusieurs interviews de survivants et (ou)

d'exhumations. En outre, les appariements de ces diverses sources de données produisaient des contradictions supplémentaires entre les dates.

Par conséquent, un descripteur a été attribué à chacune des dates du décès d'un individu afin d'indiquer l'exactitude probable de la date en question : exacte, approximative, imprécise ou inconnue. Ceci a permis de choisir une date pour chaque individu selon un système logique. Grossièrement, la date de décès attribuée à un individu était soit celle dont la précision était la plus grande ou la date la plus antérieure lorsque deux dates avaient le même niveau d'exactitude. Les hypothèses sous-tendant ce système sont que 1) les cadavres n'avaient pas été enterrés immédiatement, donc que les personnes interviewées ne les avaient pas nécessairement vus le jour du décès et 2) les dates plus exactes étaient le plus probablement celles recueillies au moment le plus rapproché de la date réelle du décès. Cependant, aucune date de décès n'a pu être attribuée à 204 enregistrements individuels. Nous avons décidé d'estimer la date de décès de ces individus par une méthode hot-deck consistant à attribuer aléatoirement une date à l'individu à partir d'enregistrements donneurs pour lesquels les décès étaient les plus proches géographiquement de l'individu pour lequel la date du décès manquait. Durant cette procédure, on a sélectionné trois dates pour chaque enregistrement et on a attribué un coefficient de pondération de 1/3 à chacune.

Après avoir choisi une date pour chaque décès individuel, on a utilisé l'estimation à systèmes multiples pour estimer le nombre de décès pour des périodes de deux jours par recoupement des quatre sources de données disponibles. Le résultat principal de l'analyse est une série d'estimations par tranche de jour des nombres de migrations et de décès au Kosovo entre mars et juin 1999. Comme le montre la figure qui suit, les dates attribuées à chaque décès et à chaque migration sont essentielles à la démonstration de l'association entre les migrations et les décès. En outre, puisque les accalmies dans les décès peuvent correspondre à des événements historiques précis (p. ex., les 6 et 7 avril représentent un cessez-le-feu unilatéral de la part de l'armée yougoslave pour observer la Fête de Pâques orthodoxe), il est essentiel de comprendre l'exactitude de l'information sur les dates.

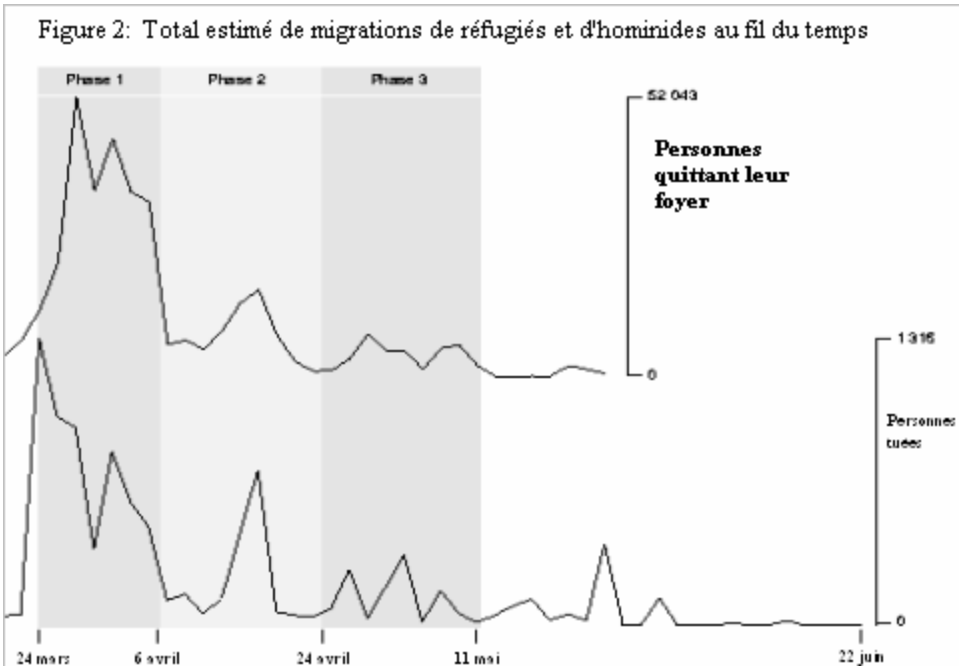

Figure provenant de la page de Ball, Betts, Scheuren, Dudukovich e t Asher (2002).

4.1 Attributions des dates pour le Kosovo

Afin de tester la sensibilité de la forme de la courbe des décès en fonction du temps au mécanisme d'attribution des dates, on a décalé les dates et exécuté de nouveau l'analyse pour déterminer la courbe résultante. On a déterminé

deux ensembles distincts de dates, à savoir un ensemble de dates « précoces » et un ensemble de dates « tardives ». Les dates utilisées étaient celles du 25^{e} et du 75^{e} percentiles pour chaque individu pour lequel on disposait de trois dates de décès ou plus et les deux dates disponibles pour les personnes pour lesquelles on disposait de deux dates de décès. Aux personnes pour lesquelles on ne possédait qu'une seule ou aucune date de décès, on a attribué une fourchette de dates par la méthode d'imputation hot-deck; pour ces enregistrement, les dates précoce et tardive correspondaient à plus ou à moins la moitié de la fourchette hot-deck. La figure qui suit montre la courbe originale en fonction du temps comparativement aux courbes pour les dates « précoce » et « tardive ». les résultats fondamentaux de l'analyse ne sont pas modifiés par la perturbation.

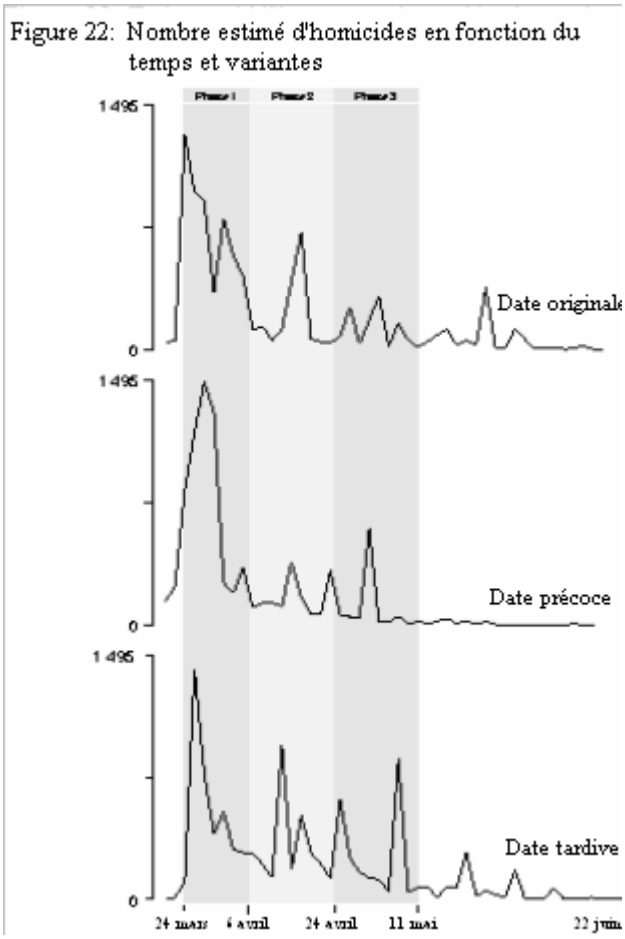

Figure provenant de la page 61 de Ball, Betts, Scheuren, Dudukovich et Asher (2002).

4.3 Avantages et problèmes

Un problème important que posent ces dates est qu'elles sont établies d'après une reconstruction rétrospective malgré leur nature intrinsèquement stochastique. Il est possible que les personnes interviewées oublient certains détails des décès dont elles ont été témoins à mesure que le temps s'écoule, ce qui modifie considérablement les résultats de l'analyse. Par conséquent, l'équipe d'analystes a dû veiller à donner tous les bénéfices du doute à l'accusé; les données qui n'étaient pas fiables ont été systématiquement exclues de l'analyse. Par exemple, aucun rapport de décès pour lequel il était impossible de déterminer un identificateur unique (p. ex. nom complet) n'a été utilisé.

Pour les décès pour lesquels existait un identificateur unique, les différences entre les dates déclarées du décès étaient problématiques. Comme nous l'avons décrit, on a utilisé un schéma logique pour déterminer une

« meilleure » date du décès, et une analyse de sensibilité donne à penser que choisir des valeurs plus extrêmes pour les dates de décès ne modifie pas significativement la courbe de décès ni les conclusions de l'analyse. Une solution éventuellement meilleure aurait consisté à utiliser des modèles formels intégrant l'incertitude concernant les dates de décès.

Cet exemple diffère des deux précédents en ce sens que les estimations des chiffres de population (ici, les dénombrements des décès) sont requis pour plusieurs points dans le temps au lieu d'un seul. L'exemple suivant représente le cas le plus puissant de données qui appuient un modèle stochastique.

5. AMERICAN COMMUNITY SURVEY

5.1 Données régionales, actualité et l'American Community Survey

Au cours des six dernières décennies, le système statistique des États-Unis s'est fondé sur le « questionnaire détaillé » du recensement décennal pour recueillir des données régionales sur le revenu, les caractéristiques du logement et d'autres sujets, outre les renseignements démographiques de base recueillis au moyen du questionnaire abrégé du recensement. Le questionnaire détaillé est distribué à un échantillon d'environ 1/6 des ménages, pour lequel le taux d'échantillonnage est plus élevé dans les régions assez peu peuplées. Un inconvénient majeur de l'utilisation du questionnaire détaillé comme source de données régionales est que la collecte des données n'a lieu qu'une fois tous les dix ans. Donc, par exemple, en 2001, les données les plus récentes disponibles avaient été recueillies lors du Recensement décennal de 1990 pour lequel l'année de référence pour le revenu était 1989 et, par conséquent, les données étaient vieilles de 12 ans, une base médiocre pour la planification et la répartition des ressources.

Nous avons envisagé deux stratégies pour obtenir des estimations plus à jour. L'une consiste à utiliser les méthodes statistiques d'estimation sur petit domaine pour mettre à jour les estimations décennales fondées sur le questionnaire détaillé du recensement à l'aide de données annuelles provenant de systèmes administratifs, comme les déclarations de revenu et l'inscription au programme de coupons alimentaires, ainsi que les données d'enquêtes courantes et à combiner les diverses sources au moyen de modèles de régression. Cette démarche a été utilisée pour produire les estimations du nombre d'enfants d'âge scolaire vivant dans la pauvreté sur lesquelles se fonde en partie la répartition du budget fédéral de l'éducation entre les autorités enseignantes locales (Citro et Kalton, 2000a, 2000b). Pour la plupart des années, les estimations calculées d'après ce programme étaient nettement plus exactes que celles reportées du recensement décennal antérieur. Cependant, l'erreur de prévision de ce modèle est plus grande que l'erreur du recensement pour l'année durant laquelle il est réalisé.

La deuxième stratégie consiste à lancer une nouvelle enquête conçue pour recueillir des données plus actuelles. L'American Community Survey (ACS), qui s'inspire du recensement continu proposé par Kish (1990, 1998) remplacerait le questionnaire détaillé du recensement décennal par une enquête réalisée de façon continue au cours de la décennie, et on ne procéderait à la collecte décennale que pour le dénombrement de population de base et les renseignements démographiques du « questionnaire abrégé » (Alexander, 2001). (L'ACS contribuerait aussi à l'établissement des chiffres de population intercensitaires, car elle dénombrerait les résidents des unités existantes et sa base de sondage inclurait les adresses des unités de logement nouvellement construites). En plus de produire des données d'une plus grande actualité, ce programme pourrait offrir des avantages opérationnels et de qualité des données significatifs en retirant la collecte des données du questionnaire détaillé de l'opération décennale surchargée et en la confiant à des employés professionnels permanents.

Quand l'ACS sera mise en oeuvre complètement, chaque échantillon mensuel sera un échantillon probabiliste national. Sur cinq ans, l'échantillon cumulatif de l'ACS sera égal en effectif à l'échantillon décennal répondant au questionnaire détaillé. Pour les grands domaines, les données d'une seule année permettront de produire des estimations suffisamment exactes, tandis que pour les petits domaines, une cumulation sur cinq années (regroupant des données au cours du temps) pourrait être nécessaire. Bien que les statistiques fondées sur une telle cumulation ne soient représentatives d'aucun point particulier dans le temps, même une cumulation sur cinq ans n'est, en moyenne, vieille que d'environ trois ans, comparativement aux données décennales qui datent, en moyenne, de plus de cinq

ans. (Au lieu d'une simple cumulation, équivalant à une moyenne mobile non pondérée, on peut utiliser une moyenne mobile pondérée qui accorde moins de poids aux années les plus éloignées, ce qui rend les données plus actuelles au prix d'une légère augmentation de la variance.) Donc, le passage à l'ACS équivaudra au remplacement d'un instantané très net par un film légèrement flou (dans les petits détails) comptant une prise de vue par année (ou, en principe, une par mois, bien qu'il soit peu probable que les données soient disponibles selon un tel calendrier).

5.2 Utilisation de données recueillies en continu pour la planification et la répartition

Il peut sembler évident que des données plus à jour représentent un meilleur fondement pour la planification — par exemple, on ne construirait pas de routes en s'appuyant sur la répartition de la population observée dix ans plus tôt si l'on disposait de données plus récentes. Même pour ce genre d'application, la transition à un nouveau calendrier de diffusion des données nécessite des modifications importantes des procédures et, élément peut-être plus critique, un changement d'état d'esprit afin d'utiliser effectivement des données qui représentent une cible temporairement en mouvement.

Le passage d'estimations décennales basées sur le questionnaire détaillé à des estimations annuelles basées sur l'ACS a certaines répercussions particulièrement intéressantes en ce qui concerne la répartition des fonds fédéraux entre les autorités locales au moyen de formules basées sur la statistique régionale, comme les taux de pauvreté ou le nombre d'enfants vivant dans la pauvreté. Dans la suite de la section, nous nous concentrons sur ces répercussions, en nous inspirant des conclusions de Zaslavsky et Schirm (2002), qui présentent des analyses et des résultats de simulation plus détaillés.

L'une des différences importantes entre les données décennales et celles recueillies en continu est que les premières représentent un échantillon temporel d'un an de la décennie, tandis que les secondes fournissent des données annuelles. Cette distinction peut influer sur la justesse des répartitions. Une seule année n'est pas nécessairement représentative de la distribution des populations d'intérêt au cours de la décennie, parce que les cycles d'emploi et l'essor et le déclin des branches d'activité dominant la scène locale ne sont pas synchronisés à l'échelle du pays. Les mesures en continu protègent contre ce genre d'effet, parce que, à long terme, le même poids cumulatif est appliqué aux données annuelles, que l'estimation pour une seule année soit fondée sur les données recueillies pour cette année-là ou sur une cumulation.

Les conséquences inattendues les plus étonnantes sont celles des interactions entre les sources des données, la méthode d'estimation et la formule de répartition des fonds. Ces composantes du processus de répartition sont intimement liées et ne peuvent pas toujours être distinguées avec précision. Par exemple, si les procédures spécifient qu'il faut fonder la répartition sur une cumulation sur trois ans, cela fait-il partie de la méthode d'estimation ou de la formule? Toutefois, certains types de dispositions souvent utilisées dans les formules de répartition afin que ces dernières soient sensibles aux variations de la fréquence des nouvelles répartitions.

L'une de ces dispositions est celle du « changement non préjudiciable », qui limite les minorations de la part attribuée à toute unité. Par exemple, la formule pourrait garantir qu'aucune autorité locale ne reçoive moins de 80 % de son allocation précédente, afin d'assurer la stabilité des activités de programmes. Une telle disposition contraint davantage la répartition lorsqu'elle est recalculée tous les dix ans que lorsqu'elle l'est annuellement; donc, une nouvelle source de données qui donne lieu à des nouvelles répartitions plus fréquentes modifiera effectivement l'effet de la disposition visant à maintenir les changements non préjudiciables. Un point plus subtil est qu'une condition visant à maintenir le changement non préjudiciable de valeur élevée (proche de 100 %) a tendance à « faire augmenter par cran » les allocations des unités pour lesquelles les estimations ont une forte variabilité d'échantillonnage : lorsque les estimations des besoins sont aléatoirement plus élevées que les valeurs réelles, les allocations augmentent et, ensuite, la disposition qui vise à ce que le changement soit non préjudiciable la maintient élevée. La variabilité d'échantillonnage peut être réduite par cumulation au prix d'une moins grande sensibilité à l'évolution des besoins); étonnamment, pour une valeur donnée de variabilité d'échantillonnage transversale, les estimations cumulées causent *moins* d'« augmentations par cran », parce que les moyennes mobiles sont plus stables au cours du temps que les estimations indépendantes ayant la même variance.

Une autre disposition courante est l'utilisation d'un seuil comme critère de financement; par exemple, une partie du budget réservée à l'aide à l'enseignement aux États-Unis est répartie uniquement entre les autorités locales en matière d'enseignement comptant au moins 6 500 enfants pauvres ou affichant un taux de pauvreté des enfants supérieur à 15 %. De nouveau, une telle provision peut interagir avec les caractéristiques d'échantillonnage des estimations. Dans le cas des estimations les plus variables, le seuil est « lissé » en ce sens que les unités dont le taux réel est à peine inférieur au seuil ont une probabilité proche de 50 % que leurs estimations soient supérieures au seuil et l'inverse se produit pour les unités dont le taux est à peine supérieur au seuil.

Bien que ce lissage d'un seuil nécessairement arbitraire ne soit pas entièrement souhaitable, l'importance du lissage, comme les effets des « changements non préjudiciables », dépend de la variation aléatoire des estimations, donc, du plan de sondage et de la méthode d'estimation. Nous pensons qu'il est indésirable que le plan de sondage ait des répercussions sur la justesse des répartitions, à part la considération générale consistant à produire pour chaque région une estimation exacte et, dans la mesure du possible, sans biais. Donc, les méthodes d'estimation et les formules devraient être conçues de façon à minimiser ce genre d'effets, qui sont notamment associés aux caractéristiques non linéaires, comme les dispositions prévoyant des changements non préjudiciables ou des seuils. L'utilisation de méthodes d'estimation et de formes linéaires a tendance à réduire ces effets, car les valeurs réelles annuelles des variables influant sur la répartition entrent alors de façon assez prévisible et symétrique dans le processus de répartition, agrégé au fil du temps.

6. CERTAINES QUESTIONS COURANTES ET IDÉES DÉTERMINANTES

Nous avons exposé ici, grâce à une série d'exemples variés, un premier effort en vue de produire un cadre uniforme de résolution des problèmes d'inférence portant sur des mesures et des systèmes multiples. Ces exemples diffèrent en plusieurs points qui illustrent la gamme de problèmes inférentiels qu'il faut résoudre. Dans le cas de l'ACS, la date est une composante explicite du plan de sondage, pour les dossiers administratifs, les dates sont inférées d'après l'information qui figure dans les fichiers et dans l'exemple du Kosovo, les inférences au sujet des dates sont un élément critique de l'analyse. Les dossiers administratifs pourraient être regroupés pour produire des données longitudinales sur les individus, tandis que les échantillons de l'ACS sont longitudinaux pour les régions, mais non pour les individus. Le modèle de la migration se concentre sur le comportement longitudinal, tandis que d'autres exemples visent à produire des inférences pour des agrégats. Néanmoins, dans chaque cas, la taille et (ou) les caractéristiques de la population sont particulières au moment où les données ont été recueillies et à la période de référence et, dans chaque cas, il est nécessaire de comprendre le problème d'inférence en formulant un cadre stochastique pour le phénomène sous-jacent.

Plus généralement, dans de telles situations, les inférences peuvent se concentrer sur un point en particulier dans le temps ou sur l'estimation de quantités ou de processus en fonction du temps, mais les données sont particulières à la période de collecte. Il est manifestement difficile de traiter ce genre de question dans le contexte des données administratives, puisque les « dates » des enregistrements peuvent varier considérablement comparativement aux dates associées au fichier de données dans son ensemble. Par exemple, il se pourrait que ce soit la date de renouvellement d'un permis de conduire et non la date de la liste des permis qui soit la plus pertinente, à moins qu'il existe des preuves de la déclaration d'une nouvelle adresse entre les dates de renouvellement. Donc, les données possèdent une dimension temporelle supplémentaire qui fait partie de l'information contenue dans chaque observation.

La nature stochastique des enquêtes, des recensements et des données administratives oblige à : a) traiter toutes les données comme des variables aléatoires, y compris les données liées aux vérifications et aux corrections, b) simplifier les hypothèses, par exemple en affectant les individus ou les événements à des périodes ou à des emplacements spécifiques et c) réaliser des analyses fondées sur des modèles plutôt que sur le plan de sondage. Indépendant du traitement de ces points particuliers, lorsque nous envisageons des sources de données multiples, nous devons relier les diverses sources de données au cadre stochastique complet, ne fût-ce que pour « projeter » les données vers une date d'intérêt, par exemple le 1er avril 2000, qui est la date de référence du Recensement décennal de 2000 aux États-Unis. En outre, l'existence de dates multiples pour les événements et de résidences multiples pour

les individus crée des problèmes supplémentaires qui exigent l'utilisation de structures de données et de modèles plus complexes.

Nos exemples montrent que les nouvelles techniques de collecte de données, qu'elles soient fondées sur des dossiers administratifs ou des enquêtes par sondage, créent de nouvelles options et possibilités statistiques. Sachant que les changements sont cumulés au fil du temps, nous reconnaissons que des données d'une plus grande actualité pourraient être plus utiles. Donc, dans le cas de l'ACS, nous devrions accorder plus de poids aux données les plus récentes dans les échantillons cumulatifs successifs. Dans le cas des sources de données multiples pour les estimations du recensement, diverses possibilités s'offrent lorsque nous cherchons à trouver un équilibre entre l'exactitude des diverses sources de données pour l'estimation des paramètres du modèle stochastique et l'actualité des sources de ces données. En outre, les progrès dans le domaine des technologies de l'information permettent de garder des renseignements provisoirement détaillés sur les transactions entre individus, tant dans les dossiers administratifs que pour des opérations comme le recensement qui ont traditionnellement été analysées comme si elles avaient lieu à un point précis dans le temps. Ces nouveaux ensembles de données plus riches créeront de nouvelles demandes de modélisation statistique.

REMERCIEMENTS

Ces travaux de recherche ont été financés en partie par le U.S. Bureau of the Census grâce à un contrat avec le Research Triangle Institute qui a financé les activités à la Carnegie Mellon University et avec le National Opinion Research Center et Datametrics Research Inc. qui ont financé les activités à l'Université Harvard.

RÉFÉRENCES

Alexander, C. H. (2001), "The American Community Survey for economic analysis", Washington, DC: U.S. Bureau of the Census.

Asher, J. et S. E. Fienberg (2002), "The Administrative Records Experiment in 2000: An Application to Population Count Estimation via Triple Systems Estimation", *Proceedings of the Government Statistics Section, American Statistical Association.*

Ball, P., Betts, W., Scheuren, F., Dudukovich, J., et J. Asher (2002), *Killings and Refugee Flow in Kosovo March June 1999: A Report to the International Criminal Tribunal for the Former Yugoslavia*, Washington, DC: American Association for the Advancement of Science.

Citro, C. F. et G. Kalton, eds. (2000a), *Small-Area Income and Poverty Estimates: Evaluation of Current Methodology*, Washington, DC: National Academy Press, 2000.

Citro, C. F. et G. Kalton, eds. (2000b), *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*, Washington, DC: National Academy Press, 2000.

Cochran, W. G. (1977), *Sampling Techniques, Third Edition*, New York: Wiley.

Fienberg, S. E. (1980), "The measurement of crime victimization: Prospects for a panel analysis of a panel survey", *The Statistician*, 29, pp. 313-350.

Kish, L. (1990), "Rolling samples and censuses", *Survey Methodology*, 16, pp. 63-78.

Kish, L. (1998), "Space/time variations and rolling samples", *Journal of Official Statistics*, 14 , pp. 31-46.

Lyberg, L. et C. M. Cassel (2001), “Sample Surveys: The field”, In N. J. Smelser and P. Baltes (eds.), *International Encyclopedia of the Social & Behavioral Sciences*, Vol. 20, Oxford: Elsevier, pp. 13475-13480.

Pfeffermann, D., C. Skinner, et K. Humphreys (1998), “The estimation of gross flows in the presence of measurement error using auxiliary variables”, *Journal of the Royal Statistical Society, Series A*, 161, 13-32.

Särndal, C-E., B. Swensson et J. Wretman (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Stasny, E. A. (1990), “Symmetry in flows among reported victimization classifications with non-random nonresponse”, *Survey Methodology*, 16, pp. 305-330.

Stuart, E., et A. M. Zaslavsky (2001), “Using administrative records to predict census day residency”, Proceedings of the Joint Statistical Meetings, American Statistical Assocation, Alexandria, VA: American Statistical Association.

Stuart, E., et A. M. Zaslavsky (2002), “Using administrative records to predict census day residency”,in C. Gatsonis et al. (eds.) *Case Studies in Bayesian Statistics, Volume VI*, New York: Springer-Verlag, pp. 335-349.

Ybarra, L. M. R. et S. L. Lohr, (2002), “Estimates of repeat victimization using the National Crime Victimization Survey,” *Journal of Quantitative Criminology*, 18, pp. 1-22.

Zaslavsky, A. M. et A. L. Schirm (2002), “Interactions between Survey Estimates and Federal Funding Formulas”, *Journal of Official Statistics*, 18, pp. 371-391.