Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Contribution à la stratégie d'imputation du Recensement de 2021 au Royaume-Uni : le mode de réponse comme variable d'appariement dans une approche fondée sur des donneurs?

Steven Rogers, Lawrence Dyer, Brian Foley[1]

Résumé

Depuis juillet 2014, l'Office for National Statistics a pris l'engagement de tenir le Recensement de 2021 au Royaume-Uni essentiellement en ligne. L'imputation au niveau de la question jouera un rôle important dans l'ajustement de la base de données du Recensement de 2021. Les recherches montrent qu'Internet pourrait produire des données plus précises que la saisie sur papier et attirer des personnes affichant des caractéristiques particulières. Nous présentons ici les résultats préliminaires des recherches destinées à comprendre comment nous pourrions gérer ces caractéristiques dans une stratégie d'imputation pour le Recensement du Royaume-Uni de 2021. Selon nos constatations, l'utilisation d'une méthode d'imputation fondée sur des donneurs pourrait nécessiter d'envisager l'inclusion du mode de réponse comme variable d'appariement dans le modèle d'imputation sous-jacent.

Mots clés : Recensement, imputation, donneur, saisie par Internet, SCANCIR

1. Introduction

Le rajustement des données du recensement pour tenir compte du manque d'uniformité au niveau des questions et de la non-réponse (ce que l'on appellera à partir de maintenant imputation) représente une partie importante du traitement des données du recensement. La stratégie d'imputation du Recensement du Royaume-Uni de 2011 était fondée sur une approche axée sur des donneurs mise en œuvre au moyen du Système canadien de contrôle et d'imputation du recensement (SCANCIR) (Aldrich et Rogers, 2012; Wardman et Rogers, 2012; Wardman, Aldrich et Rogers, 2014). Les données du recensement ont été traitées à l'intérieur de 101 « delivery groups » (DG) comprenant en moyenne 241 000 ménages et 530 000 personnes, mis en œuvre à l'aide d'une série de modules, à savoir, démographie, culture, santé et marché du travail.

Dans le SCANCIR, les variables à l'intérieur d'un module, de même que d'autres données auxiliaires, ont servi comme variables d'appariement représentant le modèle d'imputation sous-jacent. Un algorithme du plus proche voisin a permis de créer un bassin de donneurs potentiels statistiquement similaires à l'enregistrement imputé, sur la base des données disponibles dans l'ensemble des variables d'appariement. À partir de ce bassin, le donneur qui fournissait une valeur imputée a été choisi de façon aléatoire à partir d'un ensemble d'actions d'imputation quasi minimale de changements (AIQMC) (Bankier, Lachance et Poirier 1999; SCANCIR, 2009). Cette stratégie fournit des estimations ponctuelles et des estimations de la variance de la distribution des données manquantes conditionnée par le modèle d'imputation sous-jacent (appelé à partir de maintenant estimations de l'imputation).

Par suite du Recensement de 2011, après un programme de recherche détaillé, le statisticien national a recommandé la tenue d'un Recensement du Royaume-Uni principalement en ligne en 2021, complété par l'utilisation de données administratives. Cela a été approuvé par le gouvernement du Royaume-Uni en juillet 2014. L'Office for National Statistics (ONS) est maintenant pleinement engagé dans l'élaboration d'un type différent de recensement en 2021.

[1] Steven Rogers, Office for National Statistics, Segensworth Road, Fareham, Angleterre, PO15 5RR, steven.rogers@ons.gsi.gov.uk; Lawrence Dyer, lawrence.dyer@ons.gsi.gov.uk; Brian Foley, brian.foley@ons.gsi.gov.uk

L'examen des répercussions que les données administratives et la saisie par Internet peuvent avoir sur l'exactitude des estimations de l'imputation sont deux facettes importantes de cet engagement.

Des recherches antérieures indiquent que la saisie par Internet peut comporter deux caractéristiques prometteuses. Elle semble produire des données qui comportent moins de non-réponses et davantage d'uniformité que celles saisies à partir d'un questionnaire sur papier (p. ex. Côté et Laroche, 2009). Elle semble aussi attirer des répondants dont les caractéristiques sont souvent associées à celles des personnes difficiles à dénombrer, comme des hommes en âge de travailler; des niveaux plus élevés de scolarité; un emploi de professionnel; un pays de naissance à l'extérieur du Royaume-Uni; une deuxième résidence à l'étranger (Ghee, 2014). Outre le moins grand nombre de données à imputer globalement, une augmentation dans les caractéristiques observées des personnes difficiles à dénombrer ne peut que réduire l'incertitude dans les estimations de population du Recensement habituellement associée aux données imputées.

En dépit de ces aspects prometteurs, des différences de distribution dignes de mention dans les caractéristiques des répondants par Internet et sur papier laissent aussi supposer que, lorsque l'on impute des données en mode mixte au moyen d'une méthodologie fondée sur des donneurs, il peut être nécessaire d'inclure le mode comme variable d'appariement dans le modèle d'imputation sous-jacent. En l'absence de spécifications, il existe un risque d'introduire un biais d'échantillonnage dans les estimations de l'imputation par suite du tirage aléatoire d'un donneur à partir d'un bassin de donneurs qui est potentiellement hétérogène. Un biais d'échantillonnage semble plus probable en présence d'une inégalité de taille de sous-échantillon, le taux plus élevé d'enregistrements précis et uniformes dans les réponses par Internet, couplé à un taux plus élevé d'enregistrements qui doivent être imputés dans les réponses sur papier, pouvant contribuer à augmenter ce risque.

L'objectif général de la recherche actuelle est d'établir la façon de gérer des données de modes mixtes et tous les risques connexes pour l'exactitude des estimations de l'imputation dans la stratégie d'imputation du Recensement de 2021 au Royaume-Uni. Nous présentons ici les premières constatations de certains travaux préliminaires fondés sur les données du Recensement du Royaume-Uni de 2011, qui visent à évaluer les différences entre les estimations de l'imputation à partir d'une stratégie d'imputation fondée sur des donneurs qui comprend le mode comme variable d'appariement distinct et d'une autre qui ne le comprend pas. Les analyses mettent l'accent sur le DG513, qui comprend quatre administrations locales : Westminster, Kensington et Chelsea, Camden, et Londres. La saisie des données dans cette région urbaine est réputée difficile, celle-ci figurant dans la catégorie la plus élevée de l'indice de difficulté de dénombrement de l'ONS.

Les analyses portent sur les données des ménages comprenant de 1 à 6 personnes. Elles sont aussi axées sur les variables du Recensement du Royaume-Uni de 2011 imputées dans le module des données démographiques : âge; sexe; état matrimonial; activité la semaine dernière; indicateur de deuxième adresse (Royaume-Uni, autre ou aucune); et indicateur du pays de naissance (Royaume-Uni et République d'Irlande, autre ou aucun). Lorsque l'on explore les différences entre les caractéristiques des répondants sur Internet et sur papier, quelques variables additionnelles utilisées dans d'autres recherches, comme le niveau le plus élevé de qualification, l'origine ethnique et l'industrie, sont aussi comprises.

2. Résultats et discussion

2.1 Comparaison des réponses par Internet et sur papier

Le tableau 2.1-1 présente les faits saillants des différences de distribution des caractéristiques des personnes dans le DG513 qui répondent par Internet, comparativement à celles qui répondent sur papier. Toutes les catégories à l'intérieur d'une variable sont indiquées dans les cas où il existe une différence absolue d'au moins 3 % entre les modes de réponse. Les différences positives dans le tableau indiquent des proportions plus élevées dans les données par Internet pour cette catégorie.

Tableau 2.1-1
Faits saillants des différences de distribution dans les caractéristiques des répondants par Internet et sur papier

		Papier		Internet		Différence
		(n)	(%)	(n)	(%)	(%)
Groupe d'âge	25 à 29	37,3k	9,87	18,2k	13,90	+4,03
	30 à 34	35,6k	9,42	18,5k	14,09	+4,67
	Plus de 79	14,7k	3,89	1,1k	0,80	-3,09
Sexe	Homme	178,3k	46,99	67,3k	51,11	+4,12
	Femme	201,2k	53,01	64,4k	48,89	-4,12
État matrimonial	Célibataire	207,8k	55,62	81,1k	61,58	+5,96
Activité la semaine dernière	Retraité	40,3k	11,13	5,4k	4,08	-7,05
	Au travail	179,6k	49,65	40,4k	53,62	+3,97
	Étudiant	22,5k	6,23	12,5k	9,55	+3,32
Deuxième adresse	Autre (à l'extérieur du R.-U.)	28,5k	7,80	17,0k	12,94	+5,14
Pays de naissance	R.-U. et République d'Irlande	209,7k	56,13	57,5k	43,77	-12,36
	Autre	161,1k	43,13	104k	54,34	+11,21
Industrie	Finances	27,5k	8,13	15,7k	12,24	+4,11
Origine ethnique	Anglais	158,0k	44,86	40,0k	32,23	-12,63
	Autre origine de race blanche[1]	78,8k	22,37	37,8k	30,50	+8,13
Niveau de qualification le plus élevé	Aucun	40,1k	11,02	8,0k	6,06	-4,96
	Niveau 4+[2]	148k	40,66	61,7k	46,94	+6,28

1 Ni anglais, irlandais, gitan ou voyageur irlandais.
2 Grade (B.A., B.Sc.), grade plus élevé (M.A, Ph. D., PGCE), NVQ niveaux 4-5, HNC, HND, diplôme de niveau supérieur RSA, niveau supérieur du BTEC, grade de base (NI), qualifications professionnelles (enseignement, sciences infirmières, comptabilité).

Les différences observées entre les réponses par Internet et sur papier pour le DG513 correspondent à celles observées dans d'autres recherches. Les données par Internet comprenaient des proportions plus élevées de personnes de 25 à 34 ans; de sexe masculin; célibataires; au travail ou aux études; avec un niveau élevé de scolarité. Il y avait aussi des proportions plus élevées de personnes déclarant être de race blanche, mais pas Anglais ou Irlandais; nées à l'extérieur du Royaume-Uni; et ayant une deuxième adresse à l'extérieur du Royaume-Uni. Ces différences représentent les conditions qui peuvent mener à un risque de biais dans les estimations de l'imputation.

Le tableau 2.1-2 montre une ventilation détaillée de la distribution des ménages dans le DG513, par rapport à l'objectif d'imputation des données manquantes et incohérentes.

Tableau 2.1-2
Ménages dans le DG513 : Chiffres et principales comparaisons proportionnelles

Destinés à l'imputation (n)				**Comparaisons de ligne (%)**	
	Papier	Internet	Totaux	Papier	Internet
Donneurs précis	135 717	59 192	194 909	69,63	30,37
Données manquantes et incohérentes	52 807	1 922	54 729	96,49	3,51
Totaux	188 524	61 114	249 638	75,52	24,48

Proportion d'enregistrements précis (%)		***Autre (n)***	
Papier	Internet	*Non-donneur précis*	*Papier : 39*
71,99	96,86		*Internet : 16*
		Mode mixte	*1 195*

Il convient d'abord de souligner que les données pour 1 195 ménages comprenaient des réponses fournies au moyen des deux modes. Aux fins de la présente étude, ces enregistrements ont été laissés de côté dans les analyses subséquentes. Parmi les enregistrements destinés à l'imputation, 24,48 % étaient des réponses par Internet. Conformément aux constatations d'autres recherches, 96,86 % de toutes les réponses par Internet étaient précises et uniformes comparativement à seulement 71,99 % de toutes les réponses sur papier. À noter que 96,49 % de tous les enregistrements nécessitant une imputation correspondaient à des réponses sur papier, tandis que 30,37 % de tous les donneurs potentiels correspondaient à des réponses par Internet. Des inégalités proportionnelles comme celles-là représentent des cas qui pourraient mener à un risque accru de biais dans les estimations de l'imputation.

2.2 Différences dans les estimations de l'imputation

Une comparaison des estimations de l'imputation à partir d'une stratégie d'imputation fondée sur des donneurs qui comprend le mode comme variable d'appariement distincte, comparativement à une qui ne le comprend pas, a été obtenue grâce à une expérience simple. Le module démographique du DG513 a été imputé dans le SCANCIR selon deux conditions. Dans la condition contrôlée, les donneurs pouvaient uniquement être sélectionnés s'ils avaient répondu au moyen du même mode que celui de l'enregistrement imputé. En l'absence de condition, tous les donneurs pouvaient être sélectionnés, peu importe le mode.

Les différences dans les estimations de l'imputation peuvent être mesurées en comparant les distributions des données imputées obtenues à partir de chacune des deux stratégies d'imputation. Toutefois, afin de veiller à ce que les différences possibles ne soient pas simplement attribuables à la variance d'imputation, les données ont été imputées 10 fois pour chaque condition, et les comparaisons ont été fondées sur des estimations ponctuelles jackknife et des estimations de la variance des distributions imputées ces 10 fois (voir l'annexe 1).

En général, il y avait des différences systématiques dans les estimations de l'imputation à partir de chacune des stratégies qui n'ont pu être expliquées par la variance d'imputation. Le tableau 2.2-1 montre la somme des différences absolues dans les estimations de distributions pour chacune des variables du module démographique. Il montre aussi la fourchette d'intervalles de confiance de 99 % associée à ces estimations.

Tableau 2.2-1
Différences absolues dans les distributions imputées et fourchette d'intervalles de confiance de 99 %

	|Différence| (%)	Intervalles de confiance de 99 % (à deux décimales)					
		Condition contrôlée			Sans condition		
		Min.	Moy.	Max.	Min.	Moy.	Max.
Groupe d'âge[1]	5,22	0,01	0,02	0,03	0,01	0,02	0,03
Sexe	3,31	0,04	0,04	0,04	0,08	0,08	0,08
Pays de naissance	2,54	0,01	0,02	0,03	0,01	0,02	0,03
Activité la semaine dernière	1,48	<0,01	0,01	0,02	<0,01	0,01	0,02
Deuxième adresse	1,04	0,01	0,01	0,01	0,01	0,01	0,02
État matrimonial	0,88	<0,01	0,01	0,02	<0,01	0,01	0,02

1 L'âge a été imputé pour une année et regroupé : <1; suivi par des intervalles de 5 ans; se terminant par >79

La différence globale la plus importante dans les estimations de l'imputation a touché le groupe d'âge, suivi par le sexe, le pays de naissance, l'activité la semaine dernière et la deuxième adresse. La différence globale la plus faible concernait l'état matrimonial.

Le tableau 2.2-2 fournit des faits saillants détaillés des différences de distribution entre les données imputées au moyen des conditions contrôlées et sans condition. Toutes les catégories distinctes à l'intérieur de la variable sont montrées lorsqu'il existe une différence absolue d'au moins 0,5 %. Les différences positives dans le tableau montrent que, lorsque les donneurs ont été sélectionnés sans condition, peu importe le mode, les estimations proportionnelles étaient plus élevées dans les données imputées pour cette catégorie particulière. Les différences négatives indiquent des estimations proportionnelles plus faibles.

Tableau 2.2-2
Faits saillants des différences de distribution dans les estimations de l'imputation

		Condition contrôlée (%)	Sans condition (%)	Différence (%)
Groupe d'âge	15 à 19	9,890	9,046	-0,843
	20 à 34	29,292	30,802	+1,510
	35 à 44	13,440	14,032	+0,593
Sexe	Femme	53,252	51,597	-1,655
	Homme	46,748	48,403	+1,655
Pays de naissance	R.-U. et République d'Irlande	51,062	49,956	-1,106
	ACR[1]	19,916	21,182	+1,270
Activité la semaine dernière	Au travail	36,205	36,903	+0,698
Deuxième adresse	À l'extérieur du R.-U.	4,704	5,223	+0,519

1 Aucun code requis : Étudiant ayant une adresse pendant l'année scolaire à l'extérieur du Royaume-Uni

Comparativement à la stratégie d'imputation qui incluait le mode dans le modèle d'imputation, l'absence de contrôle du mode a mené à des estimations de l'imputation plus élevées pour les 20 à 44 ans, les hommes et les personnes au travail. L'augmentation dans les estimations pour le groupe des 20 à 44 ans semble être attribuable en grande partie aux estimations plus faibles pour les 15 à 19 ans. L'absence de contrôle du mode a aussi mené à des estimations plus élevées pour les étudiants vivant à l'extérieur du Royaume-Uni pendant l'année scolaire et les personnes ayant une deuxième adresse à l'extérieur du Royaume-Uni. L'augmentation des estimations pour les étudiants vivant à l'extérieur du Royaume-Uni semblait principalement attribuable aux estimations plus faibles pour les personnes nées au Royaume-Uni ou en République d'Irlande.

Fait révélateur, les résultats montrent que les caractéristiques des personnes représentées avec des proportions plus élevées dans les données par Internet (tableau 2.1-1) étaient similaires, sinon les mêmes, que les caractéristiques estimées plus fréquemment lorsque l'imputation sans mode servait de contrainte concernant la sélection des

donneurs (tableau 2.2-2). Cela semble confirmer qu'en ne conditionnant pas l'imputation par le mode, le risque d'introduire un biais d'échantillonnage dans les estimations de l'imputation, par suite de la sélection aléatoire d'un donneur à partir d'un bassin de donneurs hétérogènes, est susceptible de se produire. Dans l'ensemble, les résultats de la présente étude laissent supposer que lorsque l'on impute des données recueillies selon des modes mixtes à partir d'une méthodologie fondée sur des donneurs, particulièrement lorsque les répondants sont libres de choisir leur mode de réponse, il peut être en fait important d'envisager l'inclusion du mode comme variable d'appariement dans le modèle d'imputation sous-jacent.

Bibliographie

Aldrich, S., L. Wardman et S. Rogers (2012). The practical implementation of the 2011 UK Census imputation methodology. Disponible en ligne : [06/01/2015] http://www.unece.org/stats/documents/2012.09.sde.html#/.

Bankier, M., M. Lachance et P. Poirer (1999). A generic implementation of the new imputation methodology. Consulté en ligne : [07/12/2011] http://www.ssc.ca/survey/documents/SSC2000_M_Bankier.pdf.

SCANCIR (2009). Guide de l'utilisateur V4.5. Équipe d'élaboration du SCANCIR. Division des méthodes d'enquêtes sociales, Statistique Canada

Côté, A-M. et D. Laroche (2009). L'Internet : Un nouveau mode de collecte au Recensement. Disponible en ligne : [06/01/2015] http://www.statcan.gc.ca/pub/11-522-x/2008000/article/10986-eng.pdf.

Ghee, K. (2014). Internet versus paper mode effects in the 2011 Census of England and Wales: Analysis of Census Quality Survey agreement rates. Disponible en ligne : [06/01/2015] http://www.unece.org/stats/documents/2014.09.census1.html#/http://www.unece.org/stats/documents/2014.09.census1.html#/.

Wardman, L., S. Aldrich et S. Rogers (2012). Item imputation of Census data in an automated production environment; advantages, disadvantages and diagnostics. Disponible en ligne : [06/01/2015] http://www.unece.org/stats/documents/2012.09.sde.html#/.

Wardman, L., S. Aldrich et S. Rogers (2014). 2011 Census item edit and imputation process. Disponible en ligne : [06/01/2015] http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/item-edit-and-imputation-process.pdf.

Annexe 1

$$\bar{\theta}_{Jack} = \frac{1}{n} \sum_{i=1}^{n} (\bar{\theta}_i) \qquad Var(\theta) = \frac{n-1}{n} \sum_{i=1}^{n} \left(\bar{\theta}_i - \bar{\theta}_{Jack}\right)^2$$