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

Vers une approche intégrant des données du recensement et des données administratives pour l'imputation au niveau de la question dans le cadre du Recensement de 2021 au Royaume-Uni

Fern Leather, Katie Sharp et Steven Rogers[1]

Résumé

En vue du Recensement de 2021 au Royaume-Uni, l'Office for National Statistics (ONS) s'est engagée à mener un programme de recherche exhaustif, afin d'explorer comment les données administratives couplées peuvent servir à appuyer les processus statistiques conventionnels. Le contrôle et l'imputation (C et I) au niveau de la question joueront un rôle important pour l'ajustement de la base de données du Recensement de 2021. Toutefois, l'incertitude associée à l'exactitude et à la qualité des données administratives disponibles jette des doutes sur l'efficacité d'une approche intégrée fondée sur des données du recensement et des données administratives en ce qui a trait au C et I. Les contraintes actuelles, qui dictent une approche anonymisée de la « clef » pour le couplage des enregistrements, afin d'assurer la confidentialité, accentuent cette incertitude. Nous fournissons les résultats préliminaires d'une étude de simulation comparant l'exactitude prédictive et l'exactitude de la distribution de la stratégie conventionnelle de C et I mise en œuvre au moyen du SCANCIR pour le Recensement de 2011 au Royaume-Uni, à celles d'une approche intégrée reposant sur des données administratives synthétiques, comme données auxiliaires, avec une erreur qui augmente de façon systématique. À cette étape initiale de la recherche, nous mettons l'accent sur l'imputation d'une année d'âge. L'objectif de l'étude est de déterminer si les données auxiliaires découlant des données administratives peuvent améliorer les estimations de l'imputation, et où se situent les différentes stratégies dans un continuum d'exactitude.

Mots-clés : recensement, imputation, données administratives, SCANCIR

1. Introduction

Le rajustement des données du recensement pour tenir compte du manque de cohérence 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 de SCANCIR (Aldrich et Rogers, 2012; Wardman et Rogers, 2012; Wardman et coll., 2014).

Dans SCANCIR, les variables à l'intérieur d'un module, de même que d'autres données auxiliaires, servent de variables d'appariement représentant le modèle d'imputation sous-jacent. Un algorithme du plus proche voisin a permis de créer un bassin de donneurs potentiellement similaires statistiquement à l'enregistrement imputé, sur la base des données disponibles dans l'ensemble des variables d'appariement. Dans ce bassin, le donneur qui fournit la valeur imputée est choisi de façon aléatoire, à partir d'un ensemble d'actions d'imputation quasi minimale de changements (NMCIA) (Bankier, 1991; 2000; Bankier et coll., 1999; Bankier et coll., 2001; SCANCIR, 2009; Winkler et Chen, 2001). Cette stratégie fournit des estimations ponctuelles et des estimations de la variance de la distribution des données manquantes conditionnées 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'ONS est maintenant pleinement engagé dans l'élaboration d'un type différent de recensement en 2021. L'examen des répercussions que

[1] Fern Leather, ONS, Segensworth Road, Fareham, Royaume-Uni, PO15 5RR (fern.leather@ons.gsi.gov.uk); Katie Sharp, ONS, Segensworth Road, Fareham, Royaume-Uni, PO15 5RR (katie.sharp@ons.gsi.gov.uk); Steven Rogers, ONS, Segensworth Road, Fareham, Royaume-Uni, PO15 5RR (steven.rogers@ons.gsi.gov.uk)

les données administratives ont sur l'exactitude des estimations de l'imputation représente par conséquent une facette importante de cet engagement.

Blum (2006) décrit un certain nombre de mécanismes grâce auxquels les dossiers administratifs peuvent contribuer au processus d'imputation. Cela comprend l'imputation cold-deck, l'amélioration de la spécification du modèle et l'assurance continue de la qualité. Le premier, l'imputation cold-deck, comporte un risque, à savoir que les données du recensement observées puissent être modifiées sur la base de valeurs non uniformes obtenues à partir d'une source administrative externe. Compte tenu de ce risque et de l'incertitude entourant à la fois la qualité des sources administratives disponibles et le mécanisme de couplage, notre recherche initiale a été axée sur l'utilisation de données administratives couplées, en tant que données auxiliaires, dans le cadre de l'approche hot-deck conventionnelle déjà utilisée par l'ONS. En comparaison, très peu de recherches sont disponibles dans ce domaine, celles-ci ayant tendance à être axées sur l'utilisation des données administratives pour éviter le recours à l'imputation hot-deck (p. ex., Farber et coll., 2005).

La recherche visait à déterminer si les données administratives couplées pouvaient améliorer l'exactitude des estimations de l'imputation pour l'âge et à donner un aperçu de l'aspect réel de l'amélioration, ainsi que de l'endroit où se situent les différentes stratégies sur un continuum d'exactitude, avec comme objectif d'établir certains principes généraux pour l'utilisation et l'évaluation des sources administratives dans le cadre du Recensement de 2021 au Royaume-Uni. On a utilisé l'âge parce qu'il s'agit d'une variable clé du recensement, dont on sait qu'elle est généralement de bonne qualité dans un certain nombre d'ensembles de données administratives, ce qui en fait une base idéale pour les recherches préliminaires.

Nous présentons ici les résultats d'une recherche fondée sur les données du Recensement de 2011 au Royaume-Uni, qui ont permis d'évaluer les différences entre les estimations de l'imputation obtenues à partir d'une stratégie d'imputation fondée sur des donneurs comprenant des données administratives synthétiques, comme données auxiliaires, sous forme d'une variable « âge administratif » couplée à chaque dossier du recensement, comparativement à l'approche conventionnelle utilisée pour le Recensement de 2011 au Royaume-Uni.

La première étape de l'analyse est axée sur l'exactitude prédictive et l'exactitude de la distribution (Chambers, 2001) des estimations de l'imputation pour un groupe « typique » du recensement dont l'âge a été perturbé de façon aléatoire pour 5 % des enregistrements et a été imputé selon les conditions suivantes : aucune donnée administrative (approche du Recensement de 2011), données administratives exactes, données administratives comportant une erreur de +/-3 ans, données administratives comportant une erreur de +/-6 ans, données administratives comportant une erreur de +/-12 ans.

On a utilisé des données démographiques pour des ménages comprenant de une à six personnes, la variable de l'âge ayant été imputée à partir du modèle d'imputation du Recensement de 2011 au Royaume-Uni mis en œuvre dans SCANCIR, qui comprenait les variables d'appariement suivantes : rapport avec la personne du ménage 1; sexe, état matrimonial; activité la semaine dernière; indicateur d'étudiant et indicateur d'adresse pendant la durée de la session; indicateur du pays de naissance; évaluation des populations difficiles à dénombrer; ainsi que variable d'appariement additionnelle de l'« âge administratif » fournie par les sources administratives synthétiques, dont le poids est égal à celui utilisé pour l'âge dans le Recensement de 2011. L'âge administratif a été établi comme étant égal à l'âge véritable du recensement pour tous les enregistrements épurés.

La deuxième étape de l'étude comprenait le calcul des répartitions des erreurs pour l'âge, à partir d'ensembles de données administratives réelles, par suite de leur appariement avec les données du Recensement de 2011 (dont on présumait qu'elles représentaient l'âge véritable), selon le prénom, le nom de famille, le sexe et le code postal. Les répartitions des erreurs ont par la suite été utilisées pour créer des variables additionnelles d'âge administratif synthétique, qui ont servi à imputer l'ensemble de données perturbées, comme ci-dessus, afin de déterminer où se trouvaient les ensembles de données administratives réelles dans le continuum des erreurs établi à l'étape 1.

2. Résultats et discussion

2.1 Exactitude prédictive

Les figures 2.1-1 à 2.1-5 illustrent les répartitions des erreurs dans l'âge imputé comparativement à l'âge véritable pour chacune des conditions expérimentales, sur la base de tous les donneurs potentiels.

Figure 2.1-1.
Condition de données administratives exactes

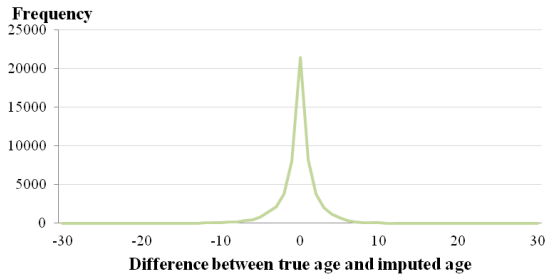


Figure 2.1-2.
Condition d'absence de données administratives

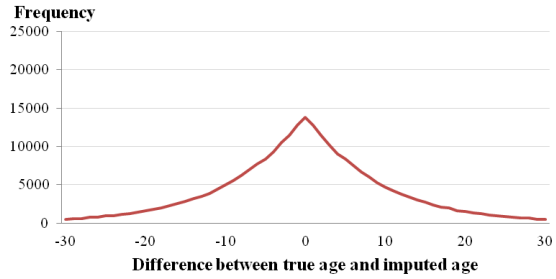


Figure 2.3-3.
Condition d'erreur de +/- 3 ans dans les données administratives

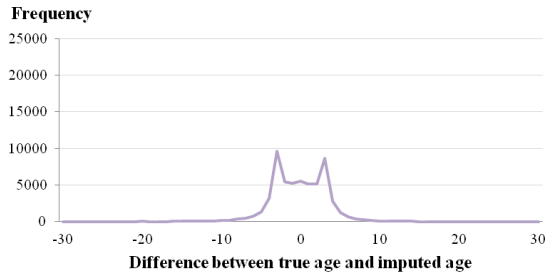


Figure 2.4-4.
Condition d'erreur de +/- 6 ans dans les données administratives

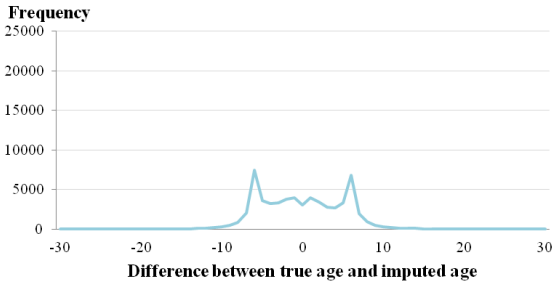
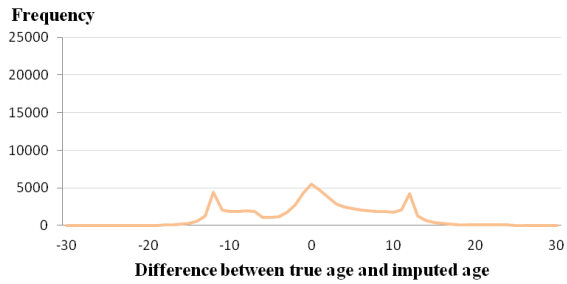


Figure 2.5-5.
Condition d'erreur de +/- 12 ans dans les données administratives



La condition de données administratives exactes a de toute évidence obtenu de meilleurs résultats que la condition d'absence de données administratives, les erreurs étant centrées autour de zéro, mais avec un écart type beaucoup plus faible (3,17, comparativement à 11,89 pour la condition d'absence de données administratives, voir le tableau 2.1-1). Les autres conditions de données administratives ont produit des répartitions non normales de l'erreur, les modes reflétant le niveau d'erreur dans la variable des données administratives, ce qui pose un problème, étant donné que cela signifie que ces conditions étaient plus susceptibles d'entraîner l'imputation d'une valeur erronée que correcte.

Tableau 2.1-1.
Statistiques sommaires pour l'erreur dans l'âge imputé comparativement à l'âge véritable pour les conditions d'absence de données administratives et de données administratives exactes

	Absence de données administratives			Données administratives exactes		
	Estimation	Intervalle de confiance à 95 %		Estimation	Intervalle de confiance à 95 %	
Moyenne	-0,15	-0,19	-0,10	-0,20	-0,22	-0,17
Écart type	11,77			3,18		
Variance	138,47			10,10		

2.2 Exactitude de la répartition

On a aussi évalué la récupération de la répartition de l'âge véritable des données perturbées. Le tableau 2.2-1 illustre les statistiques sommaires pour l'erreur dans les chiffres, par année d'âge, pour chaque condition.

Tableau 2.2-1.
Mesures de l'exactitude de la répartition pour chaque condition expérimentale

Erreur	Données administratives exactes	Absence de données administratives	Données administratives +/- 3	Données administratives +/- 6	Données administratives +/- 12
Moyenne	-0,019	-0,028	-0,086	-0,114	-0,065
IC de 95 % de la moyenne	-4,40 ; 4,36	-4,90 ; 4,84	-6,01 ; 5,84	-7,99 ; 7,76	-6,98 ; 6,85
Écart type	22,43	25,40	30,61	40,71	36,26
SSE	51 294	68 387	97 425	172 354	140 709
EMQ	22,32	25,28	30,46	40,52	36,10
IC de 95 % de l'EQM	18,57 ; 25,52	21,04 ; 28,91	22,99 36,43	30,31 ; 48,62	30,57 ; 40,88

La condition des données administratives exactes a obtenu de meilleurs résultats que la condition sans aucune donnée administrative, avec une erreur moyenne plus faible selon l'âge, une somme plus faible d'erreurs quadratiques et un écart moyen quadratique plus faible, indiquant à la fois une erreur absolue moindre pour l'ensemble de la répartition selon l'âge et une erreur moindre pour chaque âge individuel. Les autres conditions de données administratives, par ailleurs, ont difficilement permis de récupérer la répartition selon l'âge véritable à la limite de l'âge de fréquentation scolaire/l'âge actif entre 15 et 16 ans. Comme le montre le tableau 2.2-2, seulement 60 % du nombre véritable de personnes âgées de 16 ans ont été imputées pour la condition d'erreurs de +/- 6 ans, tandis que les conditions sans aucune donnée administrative et avec des données administratives exactes ont permis de récupérer avec exactitude le nombre véritable de personnes âgées de 16 ans.

Tableau 2.2-2
Fréquences postérieures à l'imputation pour les personnes âgées de 16 ans pour chaque condition expérimentale

	Valeurs véritables		Exactes		Aucune donnée administrative		Données administratives +/-3		Données administratives +/-6		Données administratives +/-12	
Âge	n	% du total	n	% du total	n	% du total	n	% du total	n	% du total	n	% du total
16	368	1,2	372	1,2	375	1,2	267	0,9	215	0,7	348	1,1

Cela semble être dû à la présence d'enregistrements comportant un âge véritable d'un côté de la limite de l'âge de fréquentation scolaire/l'âge actif et un âge administratif de l'autre côté. Les différences systématiques entre les

enregistrements de l'âge de fréquentation scolaire et de l'âge actif et la paramétrisation de SCANCIR utilisée dans le Recensement de 2011 pour maintenir ces différences ont signifié que, par exemple, il était difficile d'apparier les enregistrements comportant un âge véritable perturbé de 10 ans et un âge administratif de 16 ans avec les donneurs potentiels âgés de 16 ans, mais dans le cas des enregistrements comportant un âge véritable de 16 ans et un âge administratif de 22 ans, peu de choses sont venues entraver l'appariement avec les donneurs âgés de 22 ans, ce qui a entraîné une diminution nette du nombre de personnes âgées de 16 ans par rapport à la répartition véritable.

La condition d'erreurs +/-12 ans a obtenu de meilleurs résultats que la condition d'erreurs de +/-6 ans du point de vue de l'exactitude de la répartition, tant en ce qui a trait à l'erreur absolue pour l'ensemble de la répartition qu'à la limite de l'âge de fréquentation scolaire/l'âge actif. Cela semble venir du fait que l'erreur dans la variable de l'âge administratif était tellement importante qu'elle n'a pas pu être appariée étroitement tout en assurant l'uniformité avec les données observées, ce qui fait que l'on a accordé la priorité à l'uniformité par rapport à l'appariement selon l'âge administratif, et ainsi, qu'il était plus probable de récupérer la répartition selon l'âge véritable. Dans ce cas, par conséquent, les données observées du recensement ont été protégées efficacement par rapport à des niveaux élevés d'erreurs dans les données administratives.

2.3 Autres mesures

D'autres mesures clés de l'exactitude de l'imputation ont aussi été analysées (tableau 2.3-1). Les données administratives, peu importe leur qualité, ont entraîné une réduction de la taille du bassin de donneurs et ont limité la fourchette d'âge des donneurs potentiels à seulement ceux qui correspondaient de près à l'âge administratif. La condition de +/-12 ans a limité les donneurs dans une moindre mesure que les autres conditions, pour les raisons indiquées dans la section 2.2. À ce niveau d'erreur, il n'était simplement pas possible de trouver des donneurs pouvant être appariés étroitement selon l'âge administratif, tout en assurant l'uniformité avec les variables observées pour un grand nombre d'enregistrements.

Comme il fallait s'y attendre, le pourcentage d'enregistrements rejetés au contrôle, si l'âge administratif était substitué directement à l'âge du recensement, a augmenté en même temps que l'erreur administrative, atteignant presque 20 % pour la condition de +/-12 ans. Cela est important parce que l'on présume que les données observées du recensement sont « vraies ». Ainsi, si les données administratives ne correspondent pas avec elles, il existe un risque que les données observées puissent être modifiées sur la base des données administratives, en présence des règles de contrôle conventionnelles. Le fait que la méthode d'imputation mise en œuvre dans SCANCIR soit en mesure de protéger les données observées contre les données administratives non uniformes représente par conséquent un avantage par rapport à une méthode de substitution directe.

Tableau 2.3-1
Autres mesures de l'exactitude de l'imputation

	Données exactes	Données administratives +/-3	Données administratives +/-6	Données administratives +/-12	Aucune donnée administrative
Taille totale du bassin de donneurs	56 077	57 529	59 807	70 068	258 116
Fourchette d'âge moyen des donneurs potentiels (années)	1,0	1,1	1,4	2,5	14,3
Pourcentage d'enregistrements rejetés au contrôle si l'âge administratif remplaçait directement l'âge de recensement	0	3,8	9,1	19,4	S/O

2.4 Étape 2 : Répartitions des erreurs calculées à partir d'ensembles de données administratives réelles

Une fois les principes généraux établis et après avoir démontré de quoi avaient l'air une amélioration de l'approche sans aucune donnée administrative et le niveau d'exactitude des données administratives nécessaires pour y arriver, l'étape suivante de la recherche a été axée sur la détermination des résultats des deux ensembles de données administratives réelles dans ce contexte. À cette fin, il a d'abord fallu analyser les répartitions de l'erreur de

deux ensembles de données administratives réelles, le Registre des patients du NHS (PR) et le Système d'information des clients (CIS) du Department for Work and Pensions, couplées aux données du Recensement de 2011 selon le prénom, le nom de famille, le sexe et le code postal. Lorsque l'âge était observé dans la source administrative et dans le recensement, les ensembles de données du PR et du CIS comportaient un appariement exact de 98 % avec l'âge du recensement, ce qui montre une fiabilité élevée de la variable de l'âge.

Les répartitions des erreurs observées ont alors servi à construire une autre série d'ensembles de données administratives synthétiques, qui ont été utilisées comme données auxiliaires dans l'imputation de l'ensemble de données perturbées, de la même façon que pour les autres ensembles de données synthétiques. Cela a permis d'effectuer une analyse à l'extérieur des contraintes de l'environnement de données protégées.

Le tableau 2.4-1 énonce les répartitions de l'erreur dans l'âge imputé comparativement à l'âge véritable pour les deux conditions, en comparaison avec les conditions de données administratives exactes et d'absence de données administratives. L'exactitude prédictive était similaire à celle de la condition des données administratives exactes, ce qui était à prévoir, compte tenu de la fiabilité élevée de la variable de l'âge, et démontre clairement que les ensembles de données administratives existants peuvent permettre d'améliorer de façon substantielle la stratégie conventionnelle sans aucune donnée administrative.

Tableau 2.4-1
Statistiques sommaires pour l'erreur dans l'âge imputé comparativement à l'âge véritable pour les conditions synthétiques du CIS et du PR, par rapport aux conditions d'absence de données administratives et de données administratives exactes.

	PR			CIS			Données administratives exactes			Aucune donnée administrative		
	Estimation	IC de 95 %		Estimation	IC de 95 %		Estimation	IC de 95 %		Estimation	IC de 95 %	
Moyenne	0,20	0,17	0,23	0,21	0,18	0,24	-0,2	-0,22	-0,17	-0,15	-0,19	-0,1
Écart type	3,32	-	-	3,33	-	-	3,18	-	-	11,77	-	-
Variance	11,02	-	-	11,10	-	-	10,10	-	-	138,47	-	-

La répartition de la distribution pour les deux conditions était aussi plus élevée que pour la condition sans aucune donnée administrative, avec des chiffres d'EQM de 21,7 et 21,4 pour les conditions de PR et de CIS respectivement, comparativement à 25,3 pour la condition sans aucune donnée administrative.

2.5 Discussion générale

Les résultats montrent qu'il est possible que les données administratives améliorent de façon substantielle l'exactitude des estimations de l'imputation pour l'âge, selon certaines conditions, et que les ensembles de données administratives existants semblent être de qualité suffisamment élevée pour permettre cette amélioration. La méthodologie de changement minimal des plus proches voisins mise en œuvre dans SCANCIR a aussi permis, dans les cas où les données administratives étaient tellement erronées qu'elles ne correspondaient pas aux données observées, d'assurer le maintien de l'uniformité dans l'appariement relativement à la variable de l'âge administratif. Cela a permis de protéger efficacement les données de recensement observées contre les niveaux élevés d'erreur dans les données administratives, ce qui représentait un avantage par rapport aux autres méthodes de substitution directe.

2.6 Prochaines étapes

L'étude reposait sur un couplage parfait et une couverture des données à 100 %, afin de mettre l'accent sur les effets des erreurs dans les données administratives, ce qui ne pourrait être obtenu en situation réelle, les résultats représentant donc un scénario « idéal ». La dernière étape des travaux sera par conséquent axée sur la détermination des effets d'un couplage imparfait et d'une couverture inférieure à 100 %, afin de permettre une analyse complète coûts-avantages de la méthode, en préparation pour le Recensement de 2021.

Bibliographie

Aldrich, S., Wardman, L., and Rogers, S. (2012), "The practical implementation of the 2011 UK Census imputation methodology", *Conference of European Statisticians, Work Session on Statistical Data Editing*, UNECE.

Bankier, M. (1991), "Alternative method of doing quantitative variable imputation", Statistics Canada Memorandum.

Bankier, M., Lachance, M., and Poirier, P. (1999), "A Generic implementation of the nearest neighbour imputation method". *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 548-553.

Bankier, M. (2000), "2001 Canadian Census minimum change donor imputation methodology", *Proceedings of the UNECE Conference of European Statistics*.

Bankier, M., Poirier, P., and Lachance, M. (2001), "Efficient methodology within the Canadian Census edit and imputation system (CANCEIS)", *ASA Joint Statistical Meetings*.

Blum, O. (2006), "Evaluation of editing and imputation supported by administrative records", *Statistical Data Editing Volume 3: Impact on Data Quality*, UNECE, pp300-309.

CANCEIS (2009), "Users Guide V4.5". Social Survey Methods Division, Statistics Canada.

Chambers (2001), "National Statistics Methodological Series Report 28: Evaluation criteria for statistical editing and imputation", Office for National Statistics.

Farber, J., Wagner, D. and Resnick, D (2005), "Using Administrative Records for Imputation in the Decennial Census", *ASA Section on Survey Research Methods*.

Wardman, L., Aldrich, S., and Rogers, S. (2012), "Item imputation of Census data in an automated production environment; advantages, disadvantages and diagnostics", *Conference of European Statisticians, Work Session on Statistical Data Editing*, UNECE.

Wardman, L., Aldrich, S., and Rogers, S. (2014), "2011 Census item edit and imputation process", Office for National Statistics.

Winkler, W., & Chen, B-C. (2001). "Extending the Fellegi-Holt model of statistical data editing", *Research Report Series,* U.S. Census Bureau.