Nº 11-633-X au catalogue — Nº 006
ISSN 2371-3437
ISBN 978-0-660-07656-0

Études analytiques : méthodes et références

Imputation de codes postaux en vue de l'analyse de variables écologiques dans les cohortes longitudinales : exposition aux matières particulaires dans la base de données Cohorte santé et environnement du Recensement du Canada

par Philippe Finès, Lauren Pinault, et Michael Tjepkema

Date de diffusion : le 13 mars 2017

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Imputation de codes postaux en vue de l'analyse de variables écologiques dans les cohortes longitudinales : exposition aux matières particulaires dans la base de données Cohorte santé et environnement du Recensement du Canada

par

Philippe Finès, Lauren Pinault et Michael Tjepkema
Division de l'analyse de la santé
Statistique Canada

11-633-X Nº 006
ISSN 2371-3437
ISBN 978-0-660-07656-0

Mars 2017

Études analytiques : méthodes et références

Les documents de cette série traitent des méthodes utilisées pour produire des données qui seront employées pour effectuer des études analytiques à Statistique Canada sur l'économie, la santé et la société. Ils ont pour but de renseigner les lecteurs sur les méthodes statistiques, les normes et les définitions utilisées pour élaborer des bases de données à des fins de recherche. Tous les documents de la série ont fait l'objet d'un examen par les pairs et d'une révision institutionnelle, afin de veiller à ce qu'ils soient conformes au mandat de Statistique Canada et qu'ils respectent les normes généralement reconnues régissant les bonnes pratiques professionnelles.

Les documents peuvent être téléchargés gratuitement de **www.statcan.gc.ca**.

Table des matières

Résumé

Ce document décrit une méthode d'imputation des codes postaux manquants dans une base de données longitudinale. La base de données Cohorte santé et environnement du Recensement du Canada (CSERCan) de 1991, qui contient des renseignements sur les répondants au questionnaire détaillé du Recensement de 1991, couplée avec les fichiers des déclarations de revenus T1 pour la période allant de 1984 à 2011, est utilisée pour illustrer et valider la méthode. La cohorte contient jusqu'à 28 champs consécutifs de codes postaux de résidences, mais en raison des vides fréquents dans l'historique des codes postaux, les codes postaux manquants doivent être imputés. Pour valider la méthode d'imputation, deux expériences ont été mises au point dans lesquelles 5 % et 10 % de tous les codes postaux issus d'un sous-ensemble comportant des historiques complets ont été effacés de façon aléatoire et imputés. La proportion des écarts dans les déplacements et l'exposition moyenne était toujours plus forte dans l'expérience pour laquelle 10 % des codes postaux avaient été effacés.

Mots clés : CSERCan, cohorte de suivi du recensement; couplage de données; exposition environnementale; $PM_{2,5}$; systèmes d'information géographique; imputation; études longitudinales; pollution; codes postaux; mobilité résidentielle

1 Introduction

La base de données Cohorte santé et environnement du Recensement du Canada (CSERCan) de 1991, qui contient des renseignements sur plus de 2,5 millions de répondants au questionnaire détaillé du Recensement de 1991, a été couplée avec les données provenant des déclarations de revenus T1 pour la période allant de 1984 à 2011. En conséquence, le fichier couplé contient jusqu'à 28 champs consécutifs de codes postaux de résidences. Cependant, pour de nombreux répondants, l'historique des codes postaux est incomplet : des personnes peuvent ne pas avoir produit de déclaration de revenus ou peuvent avoir quitté le pays. En réalité, les données manquantes sur le lieu de résidence sont courantes dans les bases de données longitudinales.

Les historiques de codes postaux complets sont importants pour la recherche en hygiène de l'environnement. Les codes postaux manquants doivent être imputés afin d'attribuer des niveaux d'exposition historique aux dangers environnementaux ou aux variables écologiques à l'étude. Une imputation doit être faite de sorte que les valeurs des variables écologiques attribuées pour les années pour lesquelles des codes postaux sont manquants représentent vraisemblablement de véritables niveaux d'exposition. La méthode doit être plausible, simple et parcimonieuse, et doit produire des résultats fiables.

Cet article décrit une méthode d'imputation des codes postaux et évalue sa validité. La méthode est exposée à l'aide d'une base de données précise, CSERCan, et pour une variable écologique précise, l'exposition aux matières particulaires de 2,5 micromètres de diamètre ($PM_{2,5}$). Le concept de vide est utilisé dans tout le document. Un vide s'entend d'une série d'années *consécutives* pour lesquelles des codes postaux sont manquants.

2 Données

Le Fichier de données fiscales sommaires historiques

Le Fichier de données fiscales sommaires historiques (FDFSH) est une compilation annuelle de données provenant des déclarations de revenus représentant les personnes ayant produit une déclaration de revenus pour une année donnée. Pour la période allant de 1984 à 2011, le FDFSH fournit un historique des emplacements résidentiels des personnes et comporte 28 champs consécutifs de codes postaux (Wilkins et coll. 2008; Peters et coll. 2013). Le FDFSH a été couplé avec le fichier de la cohorte du Recensement du Canada de 1991 et la Base canadienne de données sur la mortalité (BCDM) à l'aide du numéro d'assurance sociale, ce qui a permis de créer une nouvelle base de données, CSERCan, qui fournit les historiques des codes postaux des membres de la cohorte. Ces codes postaux sont utilisés dans la recherche en hygiène de l'environnement pour attribuer des données d'exposition aux membres de la cohorte au fil du temps.

Codes postaux

Un code postal est un identificateur alphanumérique à six caractères établi et utilisé par la Société canadienne des postes aux fins du tri et de la distribution du courrier. Les caractères sont disposés selon la forme « ANA NAN », le « A » désignant un caractère alphabétique et le « N », un chiffre unique (par exemple, K1A 0T6). Les trois premiers caractères correspondent à des régions stables et précises appelées régions de tri d'acheminement (RTA); les trois derniers caractères correspondent à l'unité de distribution locale (UDL) (Statistique Canada 2014).

Le premier caractère correspond à une grande région ou à une province/territoire. Les codes dont le deuxième caractère est un zéro (0) correspondent à des régions rurales; par défaut, les codes

dont le deuxième caractère n'est pas un 0 peuvent correspondre à des régions urbaines ou suburbaines (ci-après appelées régions urbaines).

Les codes postaux ont une structure hiérarchique. Les régions dont le code comporte les deux mêmes premiers caractères se trouvent dans la grande région désignée par le premier caractère. Cette hiérarchie est reproduite avec chaque caractère supplémentaire (Statistique Canada 2014). Les codes postaux ne comprennent pas les lettres D, F, I, O, Q ou U, et les lettres W ou Z ne sont pas utilisées dans la première partie (tableau 1).

Tableau 1
Régions déterminées par le premier caractère d'un code postal

Premier caractère	Région
A	Terre-Neuve-et-Labrador
B	Nouvelle-Écosse
C	Île-du-Prince-Édouard
E	Nouveau-Brunswick
G	Est du Québec
H	Montréal métropolitain
J	Ouest du Québec
K	Est de l'Ontario
I	Centre de l'Ontario
M	Grand Toronto
N	Sud-Ouest de l'Ontario
p	Nord de l'Ontario
R	Manitoba
S	Saskatchewan
T	Alberta
V	Colombie-Britannique
X	Territoires du Nord-Ouest et Nunavut
Y	Territoire du Yukon

Source : Statistique Canada, 2014, *Fichier de conversion des codes postaux*^{MO} *plus (FCCP+) version 6C, Guide de référence – Codes postaux*^{MO} *de novembre 2014*.

3 Méthodes

Étapes préliminaires se rapportant aux codes postaux

La base de données CSERCan est utilisée pour exposer la méthode d'imputation des codes postaux, mais la méthode peut être appliquée aux autres cohortes longitudinales.

La base de données CSERCan contient 2 734 835 observations, dont 2 644 370 (celles ayant un historique valide) ont été utilisées aux fins des analyses. À l'aide du programme du Fichier de conversion des codes postaux plus (FCCP+), les codes postaux probablement non résidentiels (c'est-à-dire d'entreprises) (Statistique Canada 2014) ont été marqués et retirés de la base de données, parce qu'ils n'étaient pas susceptibles de se rapporter à une résidence privée. Parmi

les observations restantes, 1 238 825[1] (47 %) comportaient un historique complet de codes postaux, c'est-à-dire un code postal pour chaque année de suivi. Au moins un code postal était manquant entre l'année d'entrée (1984) et l'année de sortie (2011 ou l'année du décès) dans les autres observations.

Des corrections et des ajustements ont ensuite été appliquées à la base de données.

(1) **Censure :** Un membre de la cohorte qui n'était pas identifié comme étant décédé était considéré comme étant vivant jusqu'à la dernière année de suivi (2011). Dans de tels cas, si la dernière année comportant un code postal était antérieure à la dernière année de suivi :
   (a) pour les (jusqu'aux) deux premières années suivant la dernière année comportant un code postal, le dernier code postal était attribué;
   (b) pour les années subséquentes jusqu'à la dernière année de suivi, un code postal était attribué en fonction du scénario d'imputation 2b (voir ci-dessous).

(2) **Code postal au décès :** Pour les membres de la cohorte identifiés comme étant décédés (information issue de la BCDM ou du FDFSH), les règles suivantes étaient appliquées :
   (a) si la BCDM contenait un code postal, ce code postal était attribué pour cette année et était utilisé comme code postal avoisinant (« suivant ») un vide, et l'imputation reposait sur le scénario 1 (voir ci-dessous);
   (b) si la BCDM ne contenait pas de code postal, le code postal au décès était manquant. Le scénario d'imputation 2b (voir ci-dessous) était utilisé pour le dernier vide.

Imputation des codes postaux du Fichier de données fiscales sommaires historiques

L'objectif est d'utiliser l'imputation pour remplir les vides des historiques de codes postaux dans le FDFSH. Tous les codes postaux d'un vide sont imputés au cours de la même étape. Par conséquent, les valeurs imputées des codes postaux d'un même vide ne sont pas nécessairement liées entre elles, et les codes postaux nouvellement imputées ne sont pas utilisés aux fins de l'imputation.

Étant donné que la majorité des membres de la cohorte ne déménagent pas au cours d'une année donnée, les codes postaux manquants peuvent être déterminés en grande partie avec les codes postaux déclarés au cours des années avoisinantes. Par exemple, si un code postal est manquant pour une année donnée, mais que le même code postal est inscrit pour l'année précédente et l'année suivante, on peut déduire que l'adresse pendant le vide était la même que celle des années avoisinantes. Cependant, il existe toujours une probabilité non nulle que le code postal pour le vide ne soit pas imputé en fonction des codes postaux avoisinants. Par exemple, au cours d'une année, une personne peut temporairement ne pas avoir habité dans son lieu habituel de résidence.

De même, la probabilité que des codes postaux manquants puissent être imputés en fonction des codes postaux avoisinants diminue à mesure que le vide se prolonge. La méthode d'imputation tient compte d'un seuil de probabilité (p) qui varie en fonction de la durée du vide de sorte que p n'est pas 100 % et que p n'augmente pas si la durée du vide augmente. Des valeurs suggérées pour ce seuil sont présentées dans le tableau 2-1.

1. Conformément aux règles de confidentialité, les nombres qui n'étaient pas un multiple de 5 ont été arrondis de façon aléatoire au multiple de 5 les précédant ou les suivant immédiatement.

Tableau 2-1
Valeurs suggérées pour le seuil p en fonction de la durée du vide

Durée du vide en années	Seuil p
	valeur
1 ou 2	0,95
3 ou 4	0,80
5 ou plus	0,60

Des valeurs de similitude (k) et de dissimilitude (d) des codes postaux avoisinant un vide sont définies dans le tableau 2-2; reflétant la structure de codage hiérarchique des codes postaux. La valeur de similitude (k) est le nombre de caractères identiques consécutifs (à partir de la gauche) dans les codes postaux avoisinants. La valeur de dissimilitude (d) est égale à 6 moins k.

Tableau 2-2
Similitude et dissimilitude des codes postaux avoisinant le vide, à l'aide d'exemples

Code postal avant le vide	Code postal après le vide	Caractères communs avant et après le vide	Similitude (k)[1]	Dissimilitude (d)[2]
			nombre	
K1A 1A1	K1A 1A1	K1A 1A1	6	0
K1A 1A1	K1A 1A2	K1A 1A	5	1
K1A 1A1	K1A 1B1	K1A 1	4	2
K1A 1A1	K1A 2A1	K1A	3	3
K1A 1A1	K1B 1A1	K1	2	4
K1A 1A1	K2A 1A1	K	1	5
K1A 1A1	L1A 1A1	(aucun)	0	6

1. Nombre de caractères communs des codes postaux des deux côtés du vide.
2. La valeur est égale à 6 moins k.

Définition des règles et des scénarios

Deux règles sont définies à l'aide du seuil (p), de la similitude (k) et de la dissimilitude (d) :

Règle A : Le code postal manquant est imputé en fonction des codes postaux avoisinant le vide. La valeur imputée contiendra ce qui suit (de gauche à droite) : les caractères k communs aux deux côtés du vide, suivis de d fois le caractère « * ».

Règle B : Une valeur non liée aux codes postaux avoisinants est donnée au code postal manquant, c'est-à-dire la valeur « DUMMYd ».

Scénario 1 : Pour les vides ayant des codes postaux avoisinants :

> Pour chaque année du vide, un nombre aléatoire (u) est sélectionné à partir d'une répartition uniforme aléatoire $U(0,1)$. Si $u \leq p$, appliquer la **Règle A**; sinon, appliquer la **Règle B**.

Scénario 2 : Pour les vides pour lesquels il manque au moins un code postal avoisinant, la **Règle B** est toujours appliquée.

(a) S'il n'y a pas de code postal *avant* ET *après* le vide (il manque des codes postaux dans tout l'historique), « DUMMY7 » est attribué à tous les codes postaux manquants;
(b) Si seul le code postal *après* le vide est manquant, « DUMMY8 » est attribué à tous les codes postaux manquants;
(c) Si seul le code postal *avant* le vide est manquant, « DUMMY9 » est attribué à tous les codes postaux manquants.

Les règles et les scénarios sont illustrés au moyen d'exemples présentés dans l'annexe. La méthode d'imputation permet donc de remplacer un code postal manquant par ce qui suit : (1) un code postal complet; (2) une valeur qui contient les caractères d'un code postal suivis d'un certain nombre de « * »; ou (3) une valeur qui commence par « DUMMY » suivie d'un chiffre de 0 à 9.

Validation

La validation de la méthode d'imputation a pour but de déterminer si les données d'exposition calculées à l'aide des codes postaux imputés sont similaires à celles calculées à l'aide des données sur les codes postaux complets originaux. À titre d'exemple, des estimations des $PM_{2,5}$ sont calculées.

Les résidences des membres de la cohorte ont été couplées sur le plan spatial à des estimations d'une couche de surface de concentration de $PM_{2,5}$ pour toute l'Amérique du Nord continentale, calculées à partir d'un modèle qui fournit les concentrations moyennes de $PM_{2,5}$ à une résolution approximative de 1 km^2 de 2004 à 2011 (van Donkelaar et coll. 2015; Pinault et coll. 2016). Les estimations ont été rétropolées de 1998 à 2003 à l'aide de la variation interannuelle exposée dans Boys et coll. (2014). Les valeurs aberrantes de $PM_{2,5}$ supérieures à 20 microgrammes par mètre cube ($\mu g/m^3$) ont été exclues de l'analyse (moins de 1 % des membres de la cohorte au cours de n'importe quelle année) (Pinault et coll. 2016). Des données sur l'exposition à la pollution atmosphérique n'étaient pas disponibles avant 1998 (quinzième année de suivi). Si plusieurs observations pour le même code postal étaient inscrites dans la base de données sur l'exposition, l'une d'entre elles a été sélectionnée de manière aléatoire.

Étant donné que le fichier d'exposition utilisé dans cette analyse renfermait des données d'exposition pour tous les codes postaux comportant au moins les trois premiers caractères (par exemple, les codes postaux comme A0A, A0A1, A0A1A, A0A1A0), il s'ensuit que (1) tous les codes postaux établis au moyen de la Règle A et contenant au moins quatre « * » et (2) tous les codes postaux établis au moyen de la Règle B sont définis comme des codes postaux non informatifs, c'est-à-dire des codes postaux imputés pour lesquels aucune donnée d'exposition n'est disponible. Pour tous les codes postaux non informatifs, une valeur manquante est attribuée aux données d'exposition. En revanche, les codes postaux imputés se terminant par $d < 4$ « * » sont en partie informatifs. La moyenne pour tous les codes postaux commençant par les caractères de même similitude k est attribuée à l'exposition environnementale pour ces codes postaux.

Considérons un code postal manquant qui fait partie d'un vide pour lequel les codes postaux avoisinants sont présents : le scénario 1 est appliqué. Si les codes postaux avoisinants n'ont pas

de caractère commun, la Règle A est appliquée : le code postal qui en résultera sera « ****** »; ou la Règle B est appliquée : le code postal qui en résultera sera « DUMMY6 ». Ainsi, même si les deux codes postaux qui en résultent sont différents (parce qu'ils sont établis à l'aide de règles différentes), ils représentent tous deux un code postal non informatif. Autrement dit, tous les codes postaux des vides entourés de codes postaux ayant une similitude égale à 0 sont toujours non informatifs.

Une validation a été réalisée sur les 1 238 825 observations pour lesquelles les historiques de codes postaux étaient complets. Un pourcentage de ces codes postaux a été effacé de manière aléatoire, et la méthode d'imputation a été appliquée aux codes postaux manquants. Les pourcentages effacés étaient de 5 % dans la première expérience (Expérience A) et de 10 % dans la deuxième (Expérience B). Ces pourcentages sont à peu près les mêmes que le pourcentage réel des codes postaux manquants dans la base de données originale (8,1 %). Les valeurs seuils présentées dans le tableau 2-1 ont été utilisées aux fins de l'imputation. Les deux nouveaux fichiers renfermant les historiques de codes postaux imputés ont été comparés à l'ensemble de données original renfermant les historiques de codes postaux complets. Les pourcentages de codes postaux imputés à l'aide de chacune des règles et les pourcentages de codes postaux imputés correspondant aux codes postaux originaux ont été calculés.

Les résultats se rapportant aux personnes ont aussi été examinés. L'écart par personnes entre l'ensemble de données original et le nouvel ensemble de données a été utilisé pour valider la méthode d'imputation. Les mesures étaient les suivantes :

- le nombre moyen de déménagements (changements de code postal);
- le déplacement géographique moyen (en fonction des coordonnées de latitude et de longitude[2] des centroïdes du code postal); et
- l'exposition moyenne pendant l'historique (en fonction de l'exposition aux $PM_{2,5}$ selon le code postal).

Les statistiques pertinentes étaient les pourcentages des observations pour lesquelles la valeur absolue de la différence entre chacun des deux nouveaux fichiers et le fichier original atteignait ou dépassait un seuil pour les grandes valeurs. Pour le nombre moyen de déménagements, le déplacement moyen et l'exposition moyenne, le seuil a été établi à 0,1; pour le nombre de déplacements, il a été établi à 2. Ces seuils correspondaient à peu près à la queue supérieure de 5 % des répartitions des variables.

Les analyses des codes postaux ont été réalisées globalement, puis en fonction du premier caractère du code postal (qui désigne de grandes régions), puis en fonction de la désignation rurale/urbaine du deuxième caractère du code postal. Les analyses liées aux personnes ont été réalisées globalement, puis en fonction du premier caractère du premier code postal de l'historique, puis en fonction de la désignation rurale/urbaine du deuxième caractère de ce code postal.

2. Au Canada, 0,1 degré de longitude a une longueur d'environ 8 km et 0,1 degré de latitude a une longueur d'environ 4 km. Par conséquent, un carré ayant un côté de 0,1 degré a une superficie d'environ 30 km^2. Un déplacement le long de la diagonale de ce carré a une longueur d'environ 10 km.

4 Résultats

Résultats globaux

Dans l'ensemble de la base de données CSERCan comportant 2 644 370 observations,

- la durée moyenne de l'historique était d'environ 26 ans (résultats non présentés);
- environ 2,4 millions de codes postaux étaient manquants (tableau 3), ce qui représente environ 8 % des codes postaux;
- la répartition de la durée des vides était très asymétrique; 55 % des vides avaient une durée de un à deux ans (tableau 3);
- lorsque l'imputation a été effectuée, les scénarios 1 (Règle A) et 2c étaient les plus fréquents (résultats non présentés).

Tableau 3
Répartition de la durée des vides dans CSERCan après retrait des codes postaux non résidentiels

Durée du vide en années	Répartition	
	nombre	pourcentage
1	972 655	40,0
2	360 670	14,8
3	204 310	8,4
4	145 625	6,0
5	125 000	5,1
6	111 065	4,6
7	74 300	3,1
8	44 515	1,8
9	42 860	1,8
10	39 975	1,6
11	34 040	1,4
12	35 225	1,4
13	33 810	1,4
14	32 710	1,3
15	32 600	1,3
16	31 335	1,3
17	29 220	1,2
18	27 670	1,1
19	26 265	1,1
20	19 780	0,8
21	7 815	0,3
22	220	0,0
23	115	0,0
24	75	0,0
25	80	0,0
26	50	0,0
27	25	0,0
Total	**2 431 995**	**100,0**

Notes : Le total peut ne pas être égal à la somme des nombres des lignes précédentes en raison des règles de confidentialité. De plus, la somme des pourcentages n'est pas égale à 100,0 % en raison de l'arrondissement.

Source : Statistique Canada, Cohorte santé et environnement du Recensement du Canada (CSERCan) de 1991 couplée au Fichier de données fiscales sommaires historiques pour 1984 à 2011.

Analyses relatives aux codes postaux

Un total de 1 735 620 vides ont été créés dans l'Expérience A, et 3 468 405 dans l'Expérience B (tableau 4-1). Dans l'Expérience A, 91 % des codes postaux manquants faisaient partie de vides qui avaient une durée de un an et 9 % faisaient partie de vides qui avaient une durée de deux ans; dans l'Expérience B, ces proportions étaient respectivement de 82 % et de 16 %. Les Règles A et B ont été appliquées dans des proportions correspondant aux paramètres du tableau 2-1. Le pourcentage global des codes postaux se correspondant parfaitement (c.-à-d., les situations dans lesquelles le code postal imputé et le code postal original étaient les mêmes) était de 76 %; les pourcentages étaient plus élevés pour les vides courts (un ou deux ans) que pour les vides longs (cinq ans ou plus). Les résultats par région et niveau géographique ont révélé les mêmes tendances (tableaux 4-2 et 4-3).

Tableau 4-1
Rendement de l'imputation dans les Expériences A et B

Expérience et durée du vide en années	Nombre de codes postaux effacés et imputés	Pourcentage de codes postaux effacés et imputés	Règle A appliquée	Règle B appliquée	Correspondances parfaites
	nombre		pourcentage		
Expérience A					
1	1 572 760	90,6	91,6	8,4	77,0
2	151 305	8,7	91,5	8,5	72,1
3	10 815	0,6	75,9	24,1	56,5
4	720	0,0	76,7	23,3	51,7
5	20	0,0	60,0	40,0	35,0
Total – Expérience A	**1 735 620**	**100,0**	**91,5**	**8,5**	**76,4**
Expérience B					
1	2 830 755	81,6	91,4	8,6	76,8
2	547 870	15,8	91,4	8,6	72,0
3	78 510	2,3	76,1	23,9	56,8
4	10 050	0,3	76,2	23,8	53,5
5	1 090	0,0	58,2	41,8	38,0
6	115	0,0	59,6	40,3	29,8
7	15	0,0	64,3	35,7	35,7
Total – Expérience B	**3 468 405**	**100,0**	**91,0**	**9,0**	**75,5**

Notes : Le total peut ne pas être égal à la somme des nombres des lignes précédentes en raison des règles de confidentialité. La somme des pourcentages peut ne pas être égale à 100,0 % en raison de l'arrondissement. Règle A : Le code postal manquant est imputé en fonction des codes postaux avoisinant le vide. Règle B : Une valeur non liée aux codes postaux avoisinants est donnée au code postal manquant.

Source : Statistique Canada, Cohorte santé et environnement du Recensement du Canada (CSERCan) de 1991 couplée au Fichier de données fiscales sommaires historiques pour 1984 à 2011.

Tableau 4-2
Pourcentage des codes postaux imputés à l'aide de la Règle A et pourcentage des codes postaux se correspondant — Expérience A

	Pourcentage des codes postaux imputés à l'aide de la Règle A							Correspondance parfaite						
	vide de 1 an	vide de 2 ans	vide de 3 ans	vide de 4 ans	vide de 5 ans	vide de 6 ans	vide de 7 ans	vide de 1 an	vide de 2 ans	vide de 3 ans	vide de 4 ans	vide de 5 ans	vide de 6 ans	vide de 7 ans
	pourcentage													
Région[1]														
Terre-Neuve-et-Labrador	91,8	91,1	73,7	85,0	‡	‡	‡	81,3	78,7	65,7	70,0	‡	‡	‡
Nouvelle-Écosse	91,6	91,4	75,0	79,2	‡	‡	‡	78,7	75,7	60,7	62,5	‡	‡	‡
Île-du-Prince-Édouard	91,8	92,1	72,7	‡	‡	‡	‡	80,7	78,1	59,1	‡	‡	‡	‡
Nouveau-Brunswick	91,7	92,5	73,4	75,0	60,0	‡	‡	76,0	71,1	50,4	33,3	60,0	‡	‡
Est du Québec	91,5	91,5	77,0	70,3	‡	‡	‡	78,2	73,3	58,5	54,7	‡	‡	‡
Montréal métropolitain	91,0	90,8	75,8	72,4	‡	‡	‡	75,3	68,7	54,2	42,1	‡	‡	‡
Ouest du Québec	91,9	92,0	76,8	79,1	‡	‡	‡	76,8	72,0	53,7	47,3	‡	‡	‡
Est de l'Ontario	91,7	91,6	76,0	77,3	‡	‡	‡	77,6	73,1	58,4	49,3	‡	‡	‡
Centre de l'Ontario	92,1	92,2	76,1	80,8	‡	‡	‡	77,5	72,6	57,0	51,3	‡	‡	‡
Grand Toronto	90,4	90,0	74,0	57,1	‡	‡	‡	76,7	72,4	55,7	35,7	‡	‡	‡
Sud-Ouest de l'Ontario	91,7	91,6	77,0	73,5	80,0	‡	‡	78,4	74,7	60,2	41,2	40,0	‡	‡
Nord de l'Ontario	91,6	91,7	76,0	75,0	‡	‡	‡	78,8	74,5	59,1	75,0	‡	‡	‡
Manitoba	91,5	91,8	78,2	80,0	‡	‡	‡	78,3	74,7	57,6	80,0	‡	‡	‡
Saskatchewan	91,4	91,4	72,3	81,8	80,0	‡	‡	79,3	75,8	57,0	54,5	‡	‡	‡
Alberta	91,6	91,3	74,9	80,7	‡	‡	‡	74,9	68,5	55,0	47,4	‡	‡	‡
Colombie-Britannique	91,9	91,8	74,2	77,4	‡	‡	‡	73,7	68,2	51,0	53,6	‡	‡	‡
Territoires du Nord-Ouest et Nunavut	91,0	91,4	69,4	‡	‡	‡	‡	73,3	65,8	41,7	‡	‡	‡	‡
Territoire du Yukon	92,3	95,0	93,3	‡	‡	‡	‡	72,7	76,1	53,3	‡	‡	‡	‡
Total	**91,6**	**91,6**	**75,7**	**76,3**	**75,0**	‡	‡	**76,9**	**72,2**	**56,2**	**50,1**	**35,0**	‡	‡
Niveau géographique[2]														
Code postal rural	91,2	91,2	77,0	80,9	69,2	‡	‡	80,9	77,9	63,8	68,5	53,8	‡	‡
Code postal urbain	91,8	91,7	75,2	74,9	85,7	‡	‡	75,6	70,2	53,6	44,8	0,0	‡	‡
Total	**91,6**	**91,6**	**75,7**	**76,3**	**75,0**	‡	‡	**76,9**	**72,2**	**56,2**	**50,1**	**35,0**	‡	‡

‡ aucun code postal n'est manquant pour cette région ou ce niveau géographique et cette durée du vide

1. La région est désignée par le premier caractère du premier code postal.

2. Le niveau géographique est désigné par le deuxième caractère du premier code postal.

Note : Règle A : Le code postal manquant est imputé en fonction des codes postaux avoisinant le vide.

Source : Statistique Canada, Cohorte santé et environnement du Recensement du Canada (CSERCan) de 1991 couplée au Fichier de données fiscales sommaires historiques pour 1984 à 2011.

Tableau 4-3
Pourcentage des codes postaux imputés à l'aide de la Règle A et pourcentage des codes postaux se correspondant — Expérience B

	Pourcentage des codes postaux imputés à l'aide de la Règle A							Correspondance parfaite						
	vide de 1 an	vide de 2 ans	vide de 3 ans	vide de 4 ans	vide de 5 ans	vide de 6 ans	vide de 7 ans	vide de 1 an	vide de 2 ans	vide de 3 ans	vide de 4 ans	vide de 5 ans	vide de 6 ans	vide de 7 ans
	pourcentage													
Région[1]														
Terre-Neuve-et-Labrador	91,4	90,9	75,7	72,2	58,8	‡	‡	80,9	77,5	62,4	57,6	58,8	‡	‡
Nouvelle-Écosse	91,5	91,3	77,8	76,3	63,6	‡	‡	78,6	74,5	60,2	57,2	54,5	‡	‡
Île-du-Prince-Édouard	91,9	91,1	75,7	67,8	40,0	‡	‡	80,8	75,7	63,2	39,0	0,0	‡	‡
Nouveau-Brunswick	91,5	91,4	76,8	88,8	70,0	‡	‡	76,0	69,8	54,5	52,3	30,0	‡	‡
Est du Québec	91,4	91,1	75,9	78,5	63,2	50,0	‡	78,1	73,6	57,8	58,2	50,6	50,0	‡
Montréal métropolitain	90,8	90,9	76,2	78,5	65,0	50,0	‡	75,1	69,9	53,7	53,7	36,7	50,0	‡
Ouest du Québec	91,8	91,6	75,6	77,5	55,3	53,3	‡	76,6	71,3	55,9	53,6	40,7	33,3	‡
Est de l'Ontario	91,7	91,5	75,6	74,1	51,9	55,6	‡	77,7	72,7	57,1	50,3	29,6	44,4	‡
Centre de l'Ontario	91,9	91,9	77,2	75,7	58,2	70,0	‡	77,4	72,4	57,7	52,1	39,7	0,0	‡
Grand Toronto	90,1	90,1	74,9	70,8	58,9	‡	‡	76,7	72,3	58,8	51,2	32,9	‡	‡
Sud-Ouest de l'Ontario	91,5	91,7	76,8	75,1	53,8	‡	71,4	78,3	74,4	60,1	56,1	41,3	‡	71,4
Nord de l'Ontario	91,4	91,3	76,0	76,8	72,0	62,5	57,1	78,6	74,6	58,6	55,3	32,0	50,0	‡
Manitoba	91,3	91,5	75,3	77,9	60,0	‡	‡	78,2	74,8	58,9	58,1	41,4	‡	‡
Saskatchewan	91,2	91,1	74,4	79,6	50,0	‡	‡	79,1	75,0	58,6	61,2	26,2	‡	‡
Alberta	91,5	91,5	76,8	76,2	55,0	69,2	‡	74,8	69,2	55,2	48,7	32,9	0,0	‡
Colombie-Britannique	91,7	91,8	76,1	75,6	56,8	66,7	‡	73,6	68,0	51,8	50,2	35,8	11,1	‡
Territoires du Nord-Ouest et Nunavut	91,0	91,6	78,2	72,7	60,0	‡	‡	72,7	68,2	55,8	27,3	0,0	‡	‡
Territoire du Yukon	92,6	92,1	77,8	70,6	‡	‡	‡	73,8	66,9	43,3	47,1	‡	‡	‡
Total	**91,4**	**91,4**	**76,1**	**76,2**	**58,2**	**59,6**	**64,3**	**76,8**	**72,0**	**56,8**	**53,5**	**38,0**	**29,8**	**35,7**
Niveau géographique[2]														
Code postal rural	91,0	90,8	76,0	76,6	53,3	52,6	57,1	80,8	77,3	62,3	60,7	34,4	39,5	0,0
Code postal urbain	91,6	91,6	76,2	76,1	59,6	63,2	71,4	75,5	70,2	54,9	51,0	39,0	25,0	71,4
Total	**91,4**	**91,4**	**76,1**	**76,2**	**58,2**	**59,6**	**64,3**	**76,8**	**72,0**	**56,8**	**53,5**	**38,0**	**29,8**	**35,7**

‡ aucun code postal n'est manquant pour cette région ou ce niveau géographique et cette durée du vide

1. La région est désignée par le premier caractère du premier code postal.

2. Le niveau géographique est désigné par le deuxième caractère du premier code postal.

Note : Règle A : Le code postal manquant est imputé en fonction des codes postaux avoisinant le vide.

Source : Statistique Canada, Cohorte santé et environnement du Recensement du Canada (CSERCan) de 1991 couplée au Fichier de données fiscales sommaires historiques pour 1984 à 2011.

Analyses relatives aux personnes

Dans les Expériences A et B, le nombre moyen de déménagements différait d'au moins 0,1 dans 1,2 % et 4,7 % des observations, respectivement; le nombre moyen de coordonnées de latitude et de longitude était différent d'au moins 2 dans 3,5 % et 11,5 % des observations; le déplacement moyen différait d'au moins 0,1 degré de latitude et de longitude dans 2,4 % et 4,5 % des observations; et l'exposition moyenne différait d'au moins 0,1 $\mu g/m^3$ dans 4,1 % et 8,1 % des observations (résultats non présentés).

Déplacement moyen

Les différences dans le déplacement moyen entre les ensembles de données expérimentaux et l'ensemble de données original ont été examinées par région (désignée par le premier caractère du premier code postal de l'historique) et par emplacement urbain ou rural (désigné par le deuxième caractère du premier code postal de l'historique). De façon générale, le pourcentage des observations pour lesquelles la différence absolue dans la distance moyenne était d'au moins 0,1 degré ne variait pas systématiquement entre les régions, à l'exception des Territoires du Nord-Ouest et du Nunavut, pour lesquels il était plus élevé (tableau 5). Les pourcentages des observations pour lesquelles la différence absolue dans la distance moyenne était d'au moins 0,1 degré étaient légèrement plus élevés pour les codes postaux ruraux que pour les codes postaux urbains.

Tableau 5
Déplacement moyen, en fonction de la région et du niveau géographique, Expérience A et Expérience B

	Expérience A		Expérience B	
	Observations	Observations pour lesquelles la différence absolue dans la distance moyenne est ≥ 0,1 degré	Observations	Observations pour lesquelles la différence absolue dans la distance moyenne est ≥ 0,1 degré
	nombre	pourcentage	nombre	pourcentage
Région[1]				
Terre-Neuve-et-Labrador	26 835	4,83	26 830	8,99
Nouvelle-Écosse	40 180	3,12	40 180	5,46
Île-du-Prince-Édouard	6 395	2,31	6 400	3,94
Nouveau-Brunswick	35 720	2,33	35 725	4,33
Est du Québec	116 715	1,75	116 720	3,28
Montréal métropolitain	100 825	1,10	100 825	2,03
Ouest du Québec	137 210	1,04	137 205	1,93
Est de l'Ontario	72 110	2,57	72 105	4,78
Centre de l'Ontario	116 930	1,40	116 935	2,64
Grand Toronto	93 710	1,33	93 710	2,46
Sud-Ouest de l'Ontario	98 330	1,24	98 330	2,44
Nord de l'Ontario	41 830	3,64	41 830	7,00
Manitoba	56 335	3,77	56 330	6,91
Saskatchewan	54 305	4,58	54 300	8,48
Alberta	112 945	4,29	112 950	7,96
Colombie-Britannique	122 230	3,70	122 230	6,71
Territoires du Nord-Ouest et Nunavut	4 610	6,75	4 585	13,24
Territoire du Yukon	1 420	3,39	1 410	6,17
Total	**1 238 635**	**2,42**	**1 238 600**	**4,48**
Niveau géographique[2]				
Code postal rural	354 180	2,83	354 150	5,22
Code postal urbain	884 455	2,25	884 450	4,18
Total	**1 238 635**	**2,42**	**1 238 600**	**4,48**

1. La région est désignée par le premier caractère du premier code postal.

2. Le niveau géographique est désigné par le deuxième caractère du premier code postal.

Note : Le total peut ne pas être égal à la somme des nombres des lignes précédentes en raison des règles de confidentialité. De plus, dans l'Expérience B, le nombre d'observations pour lesquelles des données étaient manquantes était légèrement plus élevé. Ces deux faits expliquent pourquoi le nombre d'observations est différent pour les Expériences A et B.
Source : Statistique Canada, Cohorte santé et environnement du Recensement du Canada (CSERCan) de 1991 couplée au Fichier de données fiscales sommaires historiques pour 1984 à 2011.

Exposition moyenne

Le pourcentage des observations pour lesquelles la différence absolue dans l'exposition moyenne aux $PM_{2,5}$ était d'au moins 0,1 µg/m^3 ne variait pas systématiquement entre les régions (tableau 6); il était légèrement plus élevé dans les observations dont le premier code postal indiquait une région urbaine.

Tableau 6
Exposition moyenne aux $PM_{2,5}$, en fonction de la région et du niveau géographique, Expérience A et Expérience B

	Observations pour lesquelles la différence absolue dans l'exposition moyenne aux $PM_{2,5}$ est ≥ 0,1 µg/m^3	
	Expérience A	**Expérience B**
	pourcentage	
Région[1]		
Terre-Neuve-et-Labrador	2,68	5,14
Nouvelle-Écosse	2,77	5,56
Île-du-Prince-Édouard	1,59	3,08
Nouveau-Brunswick	2,88	6,00
Est du Québec	3,66	7,21
Montréal métropolitain	5,45	10,49
Ouest du Québec	5,34	10,44
Est de l'Ontario	4,10	7,96
Centre de l'Ontario	4,79	9,31
Grand Toronto	4,74	9,20
Sud-Ouest de l'Ontario	4,63	9,12
Nord de l'Ontario	3,54	6,88
Manitoba	2,49	4,93
Saskatchewan	3,24	6,32
Alberta	4,17	8,10
Colombie-Britannique	3,42	6,75
Territoires du Nord-Ouest et Nunavut	3,59	7,41
Territoire du Yukon	3,53	7,61
Total	**4,15**	**8,11**
Niveau géographique[2]		
Codes postaux ruraux	3,39	6,70
Codes postaux urbains	4,45	8,67
Total	**4,15**	**8,11**

1. La région est désignée par le premier caractère du premier code postal.

2. Le niveau géographique est désigné par le deuxième caractère du premier code postal.

Note : $PM_{2,5}$: matières particulaires de 2,5 micromètres de diamètre.

Source : Statistique Canada, Cohorte santé et environnement du Recensement du Canada (CSERCan) de 1991 couplée au Fichier de données fiscales sommaires historiques pour 1984 à 2011.

5 Discussion

La validation a été effectuée sur un sous-ensemble de la base de données dans lequel tous les codes postaux étaient présents, mais duquel de faibles pourcentages (5 % dans l'Expérience A; 10 % dans l'Expérience B) ont été effacés puis imputés. Le pourcentage des codes postaux effacés et imputés correspond à celui observé dans le sous-ensemble.

Les résultats obtenus relativement aux codes postaux ont révélé que les Règles A et B ont été appliquées en fonction du seuil p établi *a priori* et que le pourcentage des codes postaux se correspondant parfaitement était généralement plus élevé que deux tiers pour les vides qui ne dépassaient pas deux ans.

Les résultats issus des codes postaux relativement aux personnes (nombre de déménagements, nombre de coordonnées de latitude et de longitude, déplacement fondé sur les coordonnées de latitude et de longitude) ont révélé des écarts dans 1,2 % à 3,5 % des observations pour l'Expérience A (4,5 % à 11,5 % pour l'Expérience B). Les résultats relatifs à l'exposition aux $PM_{2,5}$ ont révélé des écarts dans 4,1 % des observations (Expérience A) et 8,7 % des observations (Expérience B). Dans le contexte de l'attribution de niveaux d'exposition en hygiène de l'environnement, ces pourcentages ont été considérés comme étant satisfaisants. De plus, ils ne varient pas énormément d'une région géographique ou d'un emplacement rural/urbain à l'autre.

D'après ces résultats, la méthode d'imputation a été considérée comme étant valide. Cependant, dans les sous-ensembles utilisés aux fins de la validation, la plupart des vides créés de façon aléatoire étaient courts (un ou deux ans), tandis que dans la base de données originale, seulement 55 % des vides avaient une durée d'un ou de deux ans (tableau 3). Par conséquent, la méthode de validation a produit un pourcentage plus faible de vides longs que dans le fichier original. La raison en est que la règle aléatoire pour rendre les codes postaux manquants dans les historiques ne tient pas compte d'une certaine corrélation qui pourrait exister entre des codes postaux manquants successifs. En conséquence, le rendement de l'imputation pourrait être légèrement surestimé. Il ne s'agit pas là d'une limite de la méthode, mais plutôt d'une limite de la validation découlant de la base de données utilisée aux fins des analyses.

Néanmoins, pour toute base de données longitudinale dans laquelle les vides ne dépassent généralement pas deux ans, les codes postaux imputés seraient similaires à ceux de la base de données originale. Le fait que de longs vides existent dans la base de données CSERCan tient à sa nature. Le seul contrôle est le choix du seuil (p) établi *a priori*. Selon la situation, les analystes utilisant des bases de données comportant des vides longs pourraient appliquer des valeurs de p beaucoup plus faibles (par exemple, 0,2) lorsque le vide devient trop long, ce qui ferait augmenter le pourcentage des occurrences de la Règle B. Cela laisse entendre que la présence de vides longs (au moins quatre ans) peut rendre ardue l'imputation des codes postaux.

D'autres méthodes pourraient être utilisées pour imputer les longs vides de codes postaux,qui sont fréquents dans les bases de données. Les codes postaux des extrémités des vides pourraient d'abord être imputés, puis, en allant vers le centre, ceux qui se trouvent dans le vide pourraient être imputés à des étapes ultérieures. Toutefois, cette méthode générerait des codes postaux qui dépendent fortement des premiers imputés et pour lesquels le niveau de confiance varierait en fonction de la distance par rapport aux extrémités du vide. Une autre possibilité consisterait à imputer les codes postaux en fonction non seulement des deux codes postaux avoisinant le vide, mais aussi de ceux qui se trouvent une ou deux années plus loin. Cette méthode pourrait mettre en jeu de nombreuses hypothèses et une série de règles complexes.

6 Conclusion

Ce document décrit une méthode d'imputation de codes postaux dans une cohorte longitudinale. L'imputation reposait largement sur les codes postaux avoisinant immédiatement les vides. Une validation a été effectuée dans laquelle un pourcentage des codes postaux a été effacé de façon aléatoire d'un sous-ensemble d'historiques complets, les codes postaux effacés ont été imputés et les résultats ont été évalués. Cette méthode d'imputation de codes postaux est pleinement fonctionnelle pour la base de données Cohorte santé et environnement du Recensement du Canada et est considérée comme valide. Elle peut être adaptée à tout fichier longitudinal et à tout polluant ou à toute variable écologique.

Les programmes SAS utilisés pour mettre en œuvre les méthodes décrites dans ce document sont disponibles auprès des auteurs sur demande. Un guide de l'utilisateur est en cours de préparation.

7 Annexe

Illustration des règles d'imputation

Pour illustrer les règles d'imputation, cinq ans de suivi et sept exemples sont présentés (six premières colonnes du tableau 1 en annexe, qui montrent des exemples de codes postaux avant imputation). Pour chaque vide relevé dans chaque exemple, le tableau fournit une description du vide, indique le scénario et la manière dont les codes postaux avoisinants se comparent (les deux colonnes centrales du tableau 1 en annexe). Conformément aux règles, pour les vides qui relèvent du scénario 1, une attribution aléatoire appliquerait la Règle A ou B. Pour les exemples qui relèvent des scénarios 2a, 2b ou 2c, la Règle B serait utilisée. Les résultats (codes postaux après imputation) sont présentés dans les dernières colonnes du tableau 1 en annexe. L'exemple 4 illustre ce qui a déjà été expliqué : les trois codes postaux manquants dans le vide sont imputés simultanément et de manière indépendante. De plus, si le générateur de nombres aléatoires avait produit des nombres aléatoires différents, les codes postaux imputés dans les exemples 1, 2 (deux codes postaux imputés), 3 (troisième année seulement) et 4 (tous les trois) auraient pu être différents : la Règle A aurait pu être appliquée au lieu de la Règle B et vice versa.

Tableau 1 en annexe

Illustration des scénarios et des règles à l'aide d'exemples hypothétiques comportant cinq ans de suivi

	Codes postaux avant imputation							Codes postaux après imputation				
Exemple	**Année 1**	**Année 2**	**Année 3**	**Année 4**	**Année 5**	**Vide 1 : Description → Détermination du scénario; Comparaison des codes postaux avoisinants → Règle(s) utilisée(s)**	**Vide 2 : Description → Détermination du scénario; Comparaison des codes postaux avoisinants → Règle(s) utilisée(s)**	**Année 1**	**Année 2**	**Année 3**	**Année 4**	**Année 5**
1	K1A1A1	(vide)	K1A1A1	K1A1A1	K1A1A1	Code postal manquant pour l'année 2 = durée de 1 an; les deux codes postaux avoisinants sont présents → Scénario 1 avec p = 0,95; k = 6; d = 0 → Règle A	s.o.	K1A1A1	**K1A1A1**	K1A1A1	K1A1A1	K1A1A1
2	K1A1A1	(vide)	K1A2B2	(vide)	K1A2B2	Code postal manquant pour l'année 2 = durée de 1 an; les deux codes postaux avoisinants sont présents → Scénario 1 avec p = 0,95; k = 3; d = 3 → Règle A	Code postal manquant pour l'année 4 donnant un vide de 1 an; les deux codes postaux avoisinants sont présents → Scénario 1 avec p = 0,95; k = 6; d = 0 → Règle B	K1A1A1	**K1A*** **	K1A2B2	**DUMMY0**	K1A2B2
3	(vide)	K1A1A1	(vide)	K1A1A1	K1A1A1	Code postal manquant pour l'année 1 = durée de 1 an; aucun code postal avant le vide → Scénario 2c; s.o → Règle B	Code postal manquant pour l'année 3 = vide de 1 an; les deux codes postaux avoisinants sont présents → Scénario 1 avec p = 0,95; k = 6; d = 0 → Règle A	**DUMMY9**	K1A1A1	**K1A1A1**	K1A1A1	K1A1A1
4	K1A1A1	(vide)	(vide)	(vide)	K1A1A2	Code postal manquant pour les années 2, 3, 4 = durée de 3 ans; les deux codes postaux avoisinants sont présents → Scénario 1 avec p = 0,80; k = 5; d = 1 → Règle A pour le premier code postal manquant; Règle B pour le deuxième code postal manquant; Règle A pour le troisième code postal manquant	s.o.	K1A1A1	**K1A1A***	**DUMMY1**	**K1A1A***	K1A1A2
5	K1A1A1	K1A1A1	K1A1A1	(vide)	(vide)	Code postal manquant pour les années 4, 5 = durée de 2 ans; aucun code postal après le vide → Scénario 2b; s.o → Règle B utilisée pour les 2 codes postaux manquants	s.o.	K1A1A1	K1A1A1	K1A1A1	**DUMMY8**	**DUMMY8**
6	(vide)	(vide)	(vide)	(vide)	(vide)	Un vide de 5 ans → Scénario 2a; s.o. → Règle B utilisée pour les 5 codes postaux manquants	s.o.	**DUMMY7**	**DUMMY7**	**DUMMY7**	**DUMMY7**	**DUMMY7**
7	K1A1A1	K1A1A1	K1A1A2	K1A1A1	K1A1A1	Aucun vide → s.o.; s.o. → Aucune imputation	s.o.	K1A1A1	K1A1A1	K1A1A2	K1A1A1	K1A1A1

Notes : Les valeurs imputées sont présentées en rouge. Règle A : Le code postal manquant est imputé en fonction des codes postaux avoisinant le vide. Règle B : Une valeur non liée aux codes postaux avoisinants est donnée au code postal manquant. k : valeur de similarité — les caractères k communs aux codes postaux des deux côtés du vide; d : valeur de dissimilarité (égale à 6 moins k); DUMMY0, DUMMY1, DUMMY7, DUMMY8 ou DUMMY9 : valeurs non liées aux codes postaux avoisinants données aux codes postaux manquants; s.o. : sans objet; p : seuil de probabilité; * : emplacement d'un caractère manquant dans le code postal.

Références

Boys, B.L., R.V. Martin, A. van Donkelaar, R.J. MacDonell, N.C. Hsu, M.J. Cooper, R.M. Yantosca, Z. Lee, D.G. Streets, Q. Zhang et S.W. Wang. 2014. « Fifteen-year global time series of satellite-derived fine particulate matter ». *Environmental Science and Technology* 48 (19) : 11109 à 11118.

Peters, P.A., M. Tjepkema, R. Wilkins, P. Fines, D. L. Crouse, P.C.W. Chan et R. T Burnett. 2013. « Data Resource Profile: 1991 Canadian Census Cohort ». *American Journal of Epidemiology* 42 (5) : 1319 à 1326.

Pinault, L., M. Tjepkema, D.L. Crouse, S. Weichenthal, A. van Donkelaar, R.V. Martin, M. Brauer, H. Chen et R.T. Burnett. 2016. « Risk estimates of mortality attributed to low concentrations of ambient fine particulate matter in the Canadian Community Health Survey cohort ». *Environmental Health* 15 (1) : 18.

Statistique Canada. 2014. *Fichier de conversion des codes postaux^{MO} plus (FCCP+) version 6C, Guide de référence – Codes postaux^{MO} de novembre 2014*. Produit nº 82-F0086-XDB au catalogue de Statistique Canada. Ottawa : Statistique Canada.

van Donkelaar, A., R.V. Martin, J.D. Spurr et R.T. Burnett. 2015. « High-resolution satellite-derived $PM_{2.5}$ from optimal estimation and geographically weighted regression over North America ». *Environmental Science and Technology* 49 (17) : 10482 à 10491.

Wilkins, R., M. Tjepkema, C. Mustard et R. Choinière. 2008. « Étude canadienne de suivi de la mortalité selon le recensement, 1991 à 2001 ». *Rapports sur la santé* 19 (3) : 25 à 43. Produit nº 82-003-XPE au catalogue de Statistique Canada.