Catalogue no. 11-633-X — No. 006
ISSN 2371-3429
ISBN 978-0-660-07655-3

Analytical Studies: Methods and References

Imputing Postal Codes to Analyze Ecological Variables in Longitudinal Cohorts: Exposure to particulate matter in the Canadian Census Health and Environment Cohort Database

by Philippe Finès, Lauren Pinault, and Michael Tjepkema

Release date: March 13, 2017

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

elephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Imputing Postal Codes to Analyze Ecological Variables in Longitudinal Cohorts: Exposure to particulate matter in the Canadian Census Health and Environment Cohort Database

by

Philippe Finès, Lauren Pinault, and Michael Tjepkema
Health Analysis Division
Statistics Canada

11-633-X No. 006
ISSN 2371-3429
ISBN 978-0-660-07655-3

March 2017

Analytical Studies: Methods and References

Papers in this series provide background discussions of the methods used to develop data for economic, health, and social analytical studies at Statistics Canada. They are intended to provide readers with information on the statistical methods, standards and definitions used to develop databases for research purposes. All papers in this series have undergone peer and institutional review to ensure that they conform to Statistics Canada's mandate and adhere to generally accepted standards of good professional practice.

The papers can be downloaded free at www.statcan.gc.ca.

Table of contents

Abstract

This paper describes a method of imputing missing postal codes in a longitudinal database. The 1991 Canadian Census Health and Environment Cohort (CanCHEC), which contains information on individuals from the 1991 Census long-form questionnaire linked with T1 tax return files for the 1984-to-2011 period, is used to illustrate and validate the method. The cohort contains up to 28 consecutive fields for postal code of residence, but because of frequent gaps in postal code history, missing postal codes must be imputed. To validate the imputation method, two experiments were devised where 5% and 10% of all postal codes from a subset with full history were randomly removed and imputed. The proportion of discrepancies in displacements and mean exposure were consistently higher in the experiment with 10% removed postal codes.

Keywords: CanCHEC, census follow-up cohort; data linkage; environmental exposure; $PM_{2.5}$; geographic information systems; imputation; longitudinal studies; pollution; postal codes; residential mobility

1 Introduction

The 1991 Canadian Census Health and Environment Cohort (CanCHEC), which contains information on more than 2.5 million respondents to the 1991 Census long-form questionnaire, was linked with T1 tax return data for the 1984-to-2011 period. Thus, the linked file contains up to 28 consecutive fields for postal code of residence. However, for many respondents, the postal code history is incomplete: individuals may not have filed a tax form, or they may have left the country. In fact, missing residential information is common in longitudinal databases.

Full postal code histories are important for environmental health research. Missing postal codes must be imputed in order to assign historical exposures to environmental hazards or to the ecological variables under study. Imputation must be done so that values of ecological variables assigned in years with missing postal codes likely represent real exposures. The method must be plausible, simple and parsimonious, and should yield reliable results.

This paper describes a method of imputing postal codes and assesses its validity. The method is demonstrated using a specific database, CanCHEC, and for a specific ecological variable, exposure to particulate matter 2.5 micrometres in diameter ($PM_{2.5}$). The concept of gap is used throughout the paper. A gap is defined as a string of *consecutive* years with missing postal codes.

2 Data

The Historical Tax Summary File

The Historical Tax Summary File (HTSF) is an annual compilation of tax return data representing individuals for whom a tax declaration was filed in a given year. For the 1984-to-2011 period, the HTSF provides a history of individuals' residential locations, with 28 consecutive fields for postal code (Wilkins et al. 2008; Peters et al. 2013). The HTSF was linked to the 1991 Canadian Census Cohort and the Canadian Mortality Database (CMDB) using social insurance number, thereby creating a new database—CanCHEC—that provides cohort members' postal code histories. These postal codes are used in environmental health research to assign exposure data to cohort members over time.

Postal codes

A postal code is a six-character alpha-numeric identifier devised and maintained by the Canada Post Corporation for sorting and delivering mail. The characters are arranged in the form " ANA NAN ," where "A" represents a letter, and "N," a single-digit number (for example, K1A 0T6). The first three characters refer to a set of stable, well-defined areas, known as forward sortation areas (FSAs); the last three characters identify routes known as local delivery units (Statistics Canada 2014).

The first character represents a large region or province/territory. Postal codes with zero (0) as the second character are rural; by default, those that do not have a 0 in the second position may be suburban or urban (hereafter urban).

Postal codes have a hierarchical structure. Areas with the same first two characters are in the major region designated by the first character. This hierarchy continues with each additional character (Statistics Canada 2014). Postal codes do not include the letters D, F, I, O, Q or U, and the first position does not use the letters W or Z (Table 1).

Table 1
Regions determined by the first character of a postal code

First character	Region
A	Newfoundland and Labrador
B	Nova Scotia
C	Prince Edward Island
E	New Brunswick
G	Eastern Quebec
H	Metropolitan Montréal
J	Western Quebec
K	Eastern Ontario
L	Central Ontario
M	Metropolitan Toronto
N	Southwestern Ontario
P	Northern Ontario
R	Manitoba
S	Saskatchewan
T	Alberta
V	British Columbia
X	Northwest Territories and Nunavut
Y	Yukon Territory

Source: Statistics Canada, 2014, *Postal Code*[OM] *Conversion File Plus (PCCF+) Version 6C, Reference Guide – November 2014 Postal Codes*[OM].

3 Methods

Preliminary steps on postal codes

The CanCHEC is used to demonstrate the method of imputing postal codes, but the method can be applied to other longitudinal cohorts.

The CanCHEC database contains 2,734,835 observations, 2,644,370 of which (those with a valid history) were used for analyses. With the Postal Code Conversion File Plus (PCCF+) program, probable non-residential (business) postal codes (Statistics Canada 2014) were flagged and removed from the database because they were unlikely to relate to a private residence. Among the remaining observations, 1,238,825[1] (47%) had a full postal code history, that is, a postal code in each year of follow-up. The others were missing at least one postal code between the year of entry (1984) and the year of exit (2011 or year of death).

Corrections and adjustments were then applied to the database.

1. In accordance with confidentiality rules, numbers that are not a multiple of 5 were randomly rounded to the multiple of 5 immediately above or below.

(1) **Censoring:** A cohort member not identified as dead was considered to be alive up to the final year of follow-up (2011). In such cases, if the last year with a postal code was before the final year of follow-up:
    (a) for (up to) the first two years after the last year with a postal code, the last postal code was assigned;
    (b) for the years thereafter until the final year of follow-up, a postal code was assigned based on imputation Case 2b (see below).

(2) **Postal code at death:** For cohort members identified as dead (from the CMDB or HTSF), the following rules were applied:
    (a) if the CMDB contained a postal code, that postal code was assigned in that year and used as a surrounding postal code ("following") for a gap, and imputation was based on Case 1 (see below);
    (b) if the CMDB did not contain a postal code, the postal code at death was missing. Imputation Case 2b (see below) was used for the last gap.

Imputing Historical Tax Summary File postal codes

The objective is to use imputation to fill the gaps in postal code histories in the HTSF. All postal codes in a gap are imputed during the same step. Consequently, the imputed values of the postal codes in the same gap are not necessarily related, and newly imputed postal codes are not used for imputation.

Because a large majority of cohort members do not move in a given year, missing postal codes can be largely based on the postal codes reported in the surrounding years. For example, if a postal code for a given year is missing, but the same postal code is recorded in the years before and after, it can be inferred that the address during the gap was the same as that of the surrounding years. However, there is always a non-null probability that the postal code for the gap should not be imputed based on the surrounding postal codes. For example, in a given year, a person may have temporarily not lived in their usual place of residence.

As well, the probability that missing postal codes can be imputed based on surrounding postal codes diminishes as the gap lengthens. The imputation method takes account of a probability threshold (p) that varies by gap length such that p is not 100%, and that p does not increase if gap length increases. Suggested values for this threshold are shown in Table 2-1.

Table 2-1
Suggested values of p threshold according to gap length

Gap length in years	p threshold
	value
1 or 2	0.95
3 or 4	0.80
5 or more	0.60

Taking advantage of the hierarchical coding structure of postal codes, similarity (k) and dissimilarity (d) values of the postal codes surrounding a gap are defined as in Table 2-2. The similarity value (k) is the number of consecutive identical characters (beginning from the left) in the surrounding postal codes. The dissimilarity value (d) is equal to 6 minus k.

Table 2-2
Similarity and dissimilarity of postal codes surrounding a gap, using examples

Postal code before gap	Postal code after gap	Common characters before and after the gap	Similarity (k)[1]	Dissimilarity (d)[2]
			number	
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	(none)	0	6

1. Number of characters common to postal codes on both sides of the gap.
2. Value equals 6 minus k.

Definition of rules and cases

Two rules are defined using threshold (p), similarity (k) and dissimilarity (d):

> **Rule A:** The missing postal code is imputed based on the postal codes surrounding the gap. The imputed value will contain (from left to right): the k characters common to both sides of the gap, followed by d times character "*".
>
> **Rule B:** The missing postal code is given a value unrelated to the surrounding postal codes, that is, the value "DUMMYd."

Case 1: For gaps with surrounding postal codes:

> For each year in the gap, a random number (u) is drawn from a random uniform distribution $U(0,1)$. If $u \leq p$, apply **Rule A**; otherwise, apply **Rule B**.

Case 2: For gaps missing at least one surrounding postal code, **Rule B** is always applied.

(a) If there is no postal code *before* AND *after* the gap (all history lacks postal codes), assign all missing postal codes to "DUMMY7";
(b) If only the postal code *after* the gap is missing, assign all missing postal codes to "DUMMY8";
(c) If only the postal code *before* the gap is missing, assign all missing postal codes to "DUMMY9."

The rules and cases are illustrated with examples in the Appendix. The imputation method thus replaces a missing postal code with: (1) a complete postal code; (2) a value that contains characters from a postal code followed by a certain number of "*"; (3) or a value that starts by "DUMMY" followed by a number from 0 through 9.

Validation

The purpose of validating the imputation method is to determine if exposure data calculated using imputed postal codes are similar to those calculated using the original, complete postal code data. As an example, estimates of $PM_{2.5}$ are calculated.

Cohort member residences were spatially linked to estimates of a surface layer of $PM_{2.5}$ concentration for all mainland North America, derived from a model that provides average $PM_{2.5}$ concentrations at an approximately 1 km^2 resolution from 2004 to 2011 (van Donkelaar et al. 2015; Pinault et al. 2016). Estimates were backcast for 1998 to 2003 using the inter-annual variation in Boys et al. (2014). Outliers with $PM_{2.5}$ values greater than 20 micrograms per cubic metre ($\mu g/m^3$) were excluded from analysis (fewer than 1% of cohort members in any year) (Pinault et al. 2016). Air pollution data for exposure were not available before 1998 (fifteenth year of follow-up). When multiple observations of the same postal code were available in the exposure database, one was randomly selected.

Since the exposure file used in this analysis contained exposure data for all postal codes with at least the first 3 characters present (for example, postal codes such as A0A, A0A1, A0A1A, A0A1A0), it follows that (1) all postal codes built with Rule A and containing at least 4 "*" and (2) all postal codes built with Rule B are defined as uninformative postal codes, that is, imputed postal codes for which no exposure data are available. For all uninformative postal codes, exposure data are assigned to a missing value. By contrast, imputed postal codes ending with $d < 4$ "*" are partially informative. The environmental exposure for these postal codes is assigned as the average for postal codes starting with the same k similarity characters.

Consider a missing postal code that belongs to a gap for which the surrounding postal codes are present: Case 1 is applied. If the surrounding postal codes have no common characters, either Rule A is applied: the resulting postal code will be "******"; or Rule B is applied: the resulting postal code will be "DUMMY6." Therefore, even though the two resulting postal codes differ (because they are built with different rules), they each represent an uninformative postal code. In other words, all postal codes in gaps surrounded by postal codes with a similarity equal to 0 are always uninformative.

Validation was performed on the 1,238,825 observations with complete postal code histories. A percentage of these postal codes was randomly erased, and the imputation method was applied to the missing postal codes. The percentages erased were 5% in the first experiment (Experiment A) and 10% in the second (Experiment B). These percentages approximate the actual percentage of missing postal codes in the original database (8.1%). Imputation used the threshold values in Table 2-1. The two new files with imputed postal code histories were compared with the original dataset with complete postal code histories. The percentages of postal codes imputed according to each rule and the percentages of imputed postal codes matching the originals were calculated.

Results for individuals were also examined. The discrepancy per individual between the original and new datasets was used to validate the imputation method. The measures were:

- mean number of moves (changes in postal code);
- mean geographic displacement (based on latitude and longitude coordinates[2] of the centroids of the postal code); and
- mean exposure during history (based on exposure to $PM_{2.5}$ by postal code).

The relevant statistics were the percentages of observations for which the absolute value of the difference between each of the two new files and the original file reached or exceeded a threshold for large values. For mean number of moves, mean displacement and mean exposure, the threshold was set at 0.1; for number of displacements, it was set at 2. These thresholds roughly corresponded to the upper 5% tail of the distributions of the variables.

2. In Canada, 0.1 degree of longitude is about 8 km long, and 0.1 degree of latitude is about 4 km long. Therefore, a square with a side of 0.1 degree has an area of about 30 km^2. A displacement along the diagonal of this square is about 10 km long.

Analyses of the postal codes were performed globally, then according to the first character of the postal code (which defines large regions), then according to the rural/urban designation of the second character of the postal code. Analyses of individuals were performed globally, then according to the first character of the first postal code in the history, then according to the rural/urban designation of the second character of that code.

4 Results

Global results

In the whole CanCHEC database of 2,644,370 observations,

- the mean length of history was about 26 years (results not shown);
- about 2.4 million postal codes were missing (Table 3), which represent about 8% of postal codes;
- distribution of the lengths of gaps was highly skewed; 55% of gaps were one or two years long (Table 3);
- when imputation was performed, Cases 1 (Rule A) and 2c were the most frequent (results not shown).

Table 3
Distribution of length of gaps in CanCHEC after removing non-residential postal codes

Length of gap in years	Distribution	
	number	percent
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: The total may not be the sum of numbers in preceding lines because of confidentiality rules. Also, the percentages do not add up to 100.0% because of rounding.

Source: Statistics Canada, 1991 Canadian Census Health and Environment Cohort (CanCHEC) linked to 1984-to-2011 Historical Tax Summary File.

Analyses of postal codes

A total of 1,735,620 gaps were created in Experiment A, and 3,468,405 in Experiment B (Table 4-1. In Experiment A, 91% of missing postal codes belonged to gaps which were one year long and 9% to gaps which were two years long; in Experiment B, the figures were respectively 82% and 16%. Rules A and B were applied in proportions corresponding to the parameters in Table 2-1. The overall percentage of perfectly matched postal codes (i.e., situations where imputed and original postal code are the same) was 76%; percentages were higher for shorter (one or two years) than for longer (five or more years) gaps. Results by region and geography showed the same patterns (Tables 4-2 and 4-3).

Table 4-1
Performance of imputation for Experiments A and B

Experiment, and length of gap in years	Number of postal codes erased and imputed	Percentage of postal codes erased and imputed	Rule A applied	Rule B applied	Perfect matches
	number		percent		
Experiment 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 – Experiment A	**1,735,620**	**100.0**	**91.5**	**8.5**	**76.4**
Experiment 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 – Experiment B	**3,468,405**	**100.0**	**91.0**	**9.0**	**75.5**

Notes: The total may not be the sum of numbers in preceding lines because of confidentality reality rules. Percentages may not add up to 100.0% because of rounding. Rule A: The missing postal code is imputed based on the postal codes surrounding the gap. Rule B: The missing postal code is given a value unrelated to the surrounding postal codes.
Source: Statistics Canada, 1991 Canadian Census Health and Environment Cohort (CanCHEC) linked to 1984-to-2011 Historical Tax Summary File.

Table 4-2
Percentage of postal codes imputed with Rule A and percentage of matching postal codes — Experiment A

	Percentage of postal codes imputed with Rule A							Perfect match						
	1-year gap	2-year gap	3-year gap	4-year gap	5-year gap	6-year gap	7-year gap	1-year gap	2-year gap	3-year gap	4-year gap	5-year gap	6-year gap	7-year gap
	percent													
Region[1]														
Newfoundland and Labrador	91.8	91.1	73.7	85.0	‡	‡	‡	81.3	78.7	65.7	70.0	‡	‡	‡
Nova Scotia	91.6	91.4	75.0	79.2	‡	‡	‡	78.7	75.7	60.7	62.5	‡	‡	‡
Prince Edward Island	91.8	92.1	72.7	‡	‡	‡	‡	80.7	78.1	59.1	‡	‡	‡	‡
New Brunswick	91.7	92.5	73.4	75.0	60.0	‡	‡	76.0	71.1	50.4	33.3	60.0	‡	‡
Eastern Quebec	91.5	91.5	77.0	70.3	‡	‡	‡	78.2	73.3	58.5	54.7	‡	‡	‡
Metropolitan Montréal	91.0	90.8	75.8	72.4	‡	‡	‡	75.3	68.7	54.2	42.1	‡	‡	‡
Western Quebec	91.9	92.0	76.8	79.1	‡	‡	‡	76.8	72.0	53.7	47.3	‡	‡	‡
Eastern Ontario	91.7	91.6	76.0	77.3	‡	‡	‡	77.6	73.1	58.4	49.3	‡	‡	‡
Central Ontario	92.1	92.2	76.1	80.8	‡	‡	‡	77.5	72.6	57.0	51.3	‡	‡	‡
Metropolitan Toronto	90.4	90.0	74.0	57.1	‡	‡	‡	76.7	72.4	55.7	35.7	‡	‡	‡
Southwestern Ontario	91.7	91.6	77.0	73.5	80.0	‡	‡	78.4	74.7	60.2	41.2	40.0	‡	‡
Northern 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	‡	‡	‡
British Columbia	91.9	91.8	74.2	77.4	‡	‡	‡	73.7	68.2	51.0	53.6	‡	‡	‡
Northwest Territories and Nunavut	91.0	91.4	69.4	‡	‡	‡	‡	73.3	65.8	41.7	‡	‡	‡	‡
Yukon Territory	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**	‡	‡
Geography[2]														
Rural postal code	91.2	91.2	77.0	80.9	69.2	‡	‡	80.9	77.9	63.8	68.5	53.8	‡	‡
Urban postal code	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**	‡	‡

‡ no missing postal codes for this region or geography and this gap length

1. The region is defined by the first character of the first postal code.

2. The geography is defined by the second character of the first postal code.

Note: Rule A: The missing postal code is imputed based on the postal codes surrounding the gap.

Source: Statistics Canada, 1991 Canadian Census Health and Environment Cohort (CanCHEC) linked to 1984-to-2011 Historical Tax Summary File.

Table 4-3
Percentage of postal codes imputed with Rule A and percentage of matching postal codes — Experiment B

	Percentage of postal codes imputed with Rule A							Perfect match						
	1-year gap	2-year gap	3-year gap	4-year gap	5-year gap	6-year gap	7-year gap	1-year gap	2-year gap	3-year gap	4-year gap	5-year gap	6-year gap	7-year gap
	percent													
Region[1]														
Newfoundland and Labrador	91.4	90.9	75.7	72.2	58.8	‡	‡	80.9	77.5	62.4	57.6	58.8	‡	‡
Nova Scotia	91.5	91.3	77.8	76.3	63.6	‡	‡	78.6	74.5	60.2	57.2	54.5	‡	‡
Prince Edward Island	91.9	91.1	75.7	67.8	40.0	‡	‡	80.8	75.7	63.2	39.0	0.0	‡	‡
New Brunswick	91.5	91.4	76.8	88.8	70.0	‡	‡	76.0	69.8	54.5	52.3	30.0	‡	‡
Eastern Quebec	91.4	91.1	75.9	78.5	63.2	50.0	‡	78.1	73.6	57.8	58.2	50.6	50.0	‡
Metropolitan Montréal	90.8	90.9	76.2	78.5	65.0	50.0	‡	75.1	69.9	53.7	53.7	36.7	50.0	‡
Western Quebec	91.8	91.6	75.6	77.5	55.3	53.3	‡	76.6	71.3	55.9	53.6	40.7	33.3	‡
Eastern 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	‡
Central 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	‡
Metropolitan Toronto	90.1	90.1	74.9	70.8	58.9	‡	‡	76.7	72.3	58.8	51.2	32.9	‡	‡
Southwestern 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
Northern 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	‡
British Columbia	91.7	91.8	76.1	75.6	56.8	66.7	‡	73.6	68.0	51.8	50.2	35.8	11.1	‡
Northwest Territories and Nunavut	91.0	91.6	78.2	72.7	60.0	‡	‡	72.7	68.2	55.8	27.3	0.0	‡	‡
Yukon Territory	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**
Geography[2]														
Rural postal code	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
Urban postal code	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**

‡ no missing postal codes for this region or geography and this gap length

1. The region is defined by the first character of the first postal code.

2. The geography is defined by the second character of the first postal code.

Note: Rule A: The missing postal code is imputed based on the postal codes surrounding the gap.

Source: Statistics Canada, 1991 Canadian Census Health and Environment Cohort (CanCHEC) linked to 1984-to-2011 Historical Tax Summary File.

Analyses of individuals

In Experiments A and B, the mean number of moves differed by at least 0.1 in 1.2% and 4.7% of observations, respectively; the mean number of latitude-longitude coordinates was different by at least 2 in 3.5% and 11.5% of observations; mean displacement differed by at least 0.1 degree of latitude-longitude in 2.4% and 4.5% of observations; and mean exposure differed by at least 0.1 $\mu g/m^3$ in 4.1% and 8.1% of observations (results not shown).

Mean displacement

Differences in mean displacement between the experimental datasets and the original dataset were examined by region (defined by first character of first postal code in history), and urban versus rural location (defined by second character of first postal code in history). Generally, the percentage of observations where the absolute difference in mean distance was at least 0.1 degree did not vary systematically between regions, except for the Northwest Territories and Nunavut, where it was higher (Table 5). The percentages of observations where the absolute difference in mean distance was at least 0.1 degree were slightly higher for rural than urban postal codes.

Table 5
Mean displacement, by region and geography, Experiment A and Experiment B

	Experiment A		Experiment B	
	Observations	Observations with absolute difference in mean distance ≥ 0.1 degree	Observations	Observations with absolute difference in mean distance ≥ 0.1 degree
	number	percent	number	percent
Region[1]				
Newfoundland and Labrador	26,835	4.83	26,830	8.99
Nova Scotia	40,180	3.12	40,180	5.46
Prince Edward Island	6,395	2.31	6,400	3.94
New Brunswick	35,720	2.33	35,725	4.33
Eastern Quebec	116,715	1.75	116,720	3.28
Metropolitan Montréal	100,825	1.10	100,825	2.03
Western Quebec	137,210	1.04	137,205	1.93
Eastern Ontario	72,110	2.57	72,105	4.78
Central Ontario	116,930	1.40	116,935	2.64
Metropolitan Toronto	93,710	1.33	93,710	2.46
Southwestern Ontario	98,330	1.24	98,330	2.44
Northern 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
British Columbia	122,230	3.70	122,230	6.71
Northwest Territories and Nunavut	4,610	6.75	4,585	13.24
Yukon Territory	1,420	3.39	1,410	6.17
Total	**1,238,635**	**2.42**	**1,238,600**	**4.48**
Geography[2]				
Rural postal code	354,180	2.83	354,150	5.22
Urban postal code	884,455	2.25	884,450	4.18
Total	**1,238,635**	**2.42**	**1,238,600**	**4.48**

1. The region is defined by the first character of the first postal code.

2. The geography is defined by the second character of the first postal code.

Note: The total may not be the sum of numbers in preceding lines because of confidentiality rules. Also, in Experiment B, the number of observations with missing values was slightly higher. These two facts explain why the number of observations for Experiments A and B differ.

Source: Statistics Canada, 1991 Canadian Census Health and Environment Cohort (CanCHEC) linked to 1984-to-2011 Historical Tax Summary File.

Mean exposure

The percentage of observations where the absolute difference in mean $PM_{2.5}$ exposure was at least 0.1 µg/m^3 did not vary systematically between regions (Table 6); it was slightly higher for observations with the first postal code indicating an urban region.

Table 6
Mean $PM_{2.5}$ exposure, by region and geography, Experiment A and Experiment B

	Observations with absolute difference in mean $PM_{2.5}$ exposure ≥ 0.1 µg/m^3	
	Experiment A	**Experiment B**
	percent	
Region[1]		
Newfoundland and Labrador	2.68	5.14
Nova Scotia	2.77	5.56
Prince Edward Island	1.59	3.08
New Brunswick	2.88	6.00
Eastern Quebec	3.66	7.21
Metropolitan Montréal	5.45	10.49
Western Quebec	5.34	10.44
Eastern Ontario	4.10	7.96
Central Ontario	4.79	9.31
Metropolitan Toronto	4.74	9.20
Southwestern Ontario	4.63	9.12
Northern Ontario	3.54	6.88
Manitoba	2.49	4.93
Saskatchewan	3.24	6.32
Alberta	4.17	8.10
British Columbia	3.42	6.75
Northwest Territories and Nunavut	3.59	7.41
Yukon Territory	3.53	7.61
Total	**4.15**	**8.11**
Geography[2]		
Rural postal codes	3.39	6.70
Urban postal codes	4.45	8.67
Total	**4.15**	**8.11**

1. The region is defined by the first character of the first postal code.

2. The geography is defined by the second character of the first postal code.

Note: $PM_{2.5}$: particulate matter 2.5 micrometres in diameter.

Source: Statistics Canada, 1991 Canadian Census Health and Environment Cohort (CanCHEC) linked to 1984-to-2011 Historical Tax Summary File.

5 Discussion

Validation was conducted on a subset of the database in which all postal codes were present, but from which small percentages (5% in Experiment A; 10% in Experiment B) were erased and then imputed. The percentage of postal codes erased and imputed fit with the percentage observed in the subset.

Results for postal codes showed that Rules A and B were applied according to the a priori threshold p, and that the percentage of perfectly matching postal codes was usually greater than two-thirds for gaps that did not exceed two years.

Postal code results for individuals (number of moves, number of latitude-longitude coordinates, displacement based on latitude-longitude coordinates) revealed discrepancies in 1.2% to 3.5% of observations for Experiment A (4.5% to 11.5% for Experiment B). Results for $PM_{2.5}$ exposure revealed discrepancies in 4.1% of observations (Experiment A) and 8.7% of observations (Experiment B). In the context of assigning environmental health exposures, these percentages were considered satisfactory. Also, they did not vary drastically across geographic region or rural/urban location.

Based on these results, the imputation method was considered to be valid. However, in the subsets used for validation, most of the randomly created gaps were short (one or two years), whereas in the original database, only 55% of gaps were one or two years long (Table 3). Thus, the validation method yielded a lower percentage of long gaps than in the original file. This is because the random rule to render postal codes missing in the histories does not take account of some correlation that could exist between successive missing postal codes. Consequently, performance of the imputation could be slightly overestimated. This is not a limitation of the method, but rather, a limitation of the validation due to the database used for the analyses.

Nevertheless, for any longitudinal database in which gaps generally do not exceed two years, the imputed postal codes would be similar to those of the original database. It is the nature of CanCHEC that long gaps will occur. The only control is the choice of the a priori threshold (p). Depending on the situation, analysts using databases with long gaps could apply much lower values of p (for instance, 0.2) when the gap becomes too long, thereby increasing the percentage of occurrences of Rule B. This suggests that the presence of long gaps (at least four years) may make imputation of postal codes arduous.

Other methods could be used to impute the long postal code gaps that frequently occur in databases. Postal codes at the ends of the gaps might be imputed first, and then, working inward, those in the interior of the gap could be imputed in subsequent steps. However, this would generate postal codes that are highly dependent on the first ones imputed, and for which the level of confidence would vary with the distance to the ends of the gap. Another possibility is to impute postal codes based not only on the two surrounding the gap, but also on those that are one or two years further away. This could involve many hypotheses and a series of complex rules.

6 Conclusion

This paper describes a method to impute postal codes in a longitudinal cohort. Imputation was largely based on the postal codes immediately surrounding the gaps. Validation was conducted by randomly erasing a percentage of postal codes from a subset of full histories, imputing the postal codes that were erased, and evaluating the results. This method of imputing postal codes is fully functional for the Canadian Census Health and Environment Cohort database and is considered valid. It can be adapted for any longitudinal file and for any pollutant or ecological variable.

The SAS programs used to implement the methods described in this paper are available from the authors on request. A user guide is in preparation.

7 Appendix

Illustration of imputation rules

To illustrate the imputation rules, five years of follow-up and seven examples are presented (first six columns in Appendix Table 1, showing examples of postal codes before imputation). For each gap identified in each example, the table provides a description of the gaps, identification of the case and how the surrounding postal codes compare (the two central columns in Appendix Table 1). According to the rules, for gaps belonging to Case 1, a random assignment would apply to Rule A or B. For examples belonging to Cases 2a, 2b or 2c, Rule B would be used. Results (postal codes after imputation) are shown in the last columns of Appendix Table 1. Example 4 illustrates what has already been explained: the three missing postal codes in the gap are imputed simultaneously and independently. Also, had the random generator produced different random numbers, imputed postal codes for examples 1, 2 (both imputed postal codes), 3 (year 3 only) and 4 (the 3 of them) could have been different: Rule A could have been applied instead of Rule B, and vice versa.

Appendix Table 1
Illustration of cases and rules on hypothetical examples with 5 years of follow-up

	Postal codes before imputation					Gap 1: Description → Identification of case; Comparison of surrounding postal codes → Rule(s) used	Gap 2: Description → Identification of case; Comparison of surrounding postal codes → Rule(s) used	Postal codes after imputation				
Example	Year 1	Year 2	Year 3	Year 4	Year 5			Year 1	Year 2	Year 3	Year 4	Year 5
1	K1A1A1	(empty)	K1A1A1	K1A1A1	K1A1A1	Missing postal code in year 2 = 1-year length; both surrounding postal codes present → Case 1 with p = 0.95; k = 6; d = 0 → Rule A	n/a	K1A1A1	**K1A1A1**	K1A1A1	K1A1A1	K1A1A1
2	K1A1A1	(empty)	K1A2B2	(empty)	K1A2B2	Missing postal code in year 2 = 1-year length; both surrounding postal codes present → Case 1 with p = 0.95; k = 3; d = 3 → Rule A	Missing postal code in year 4 makes 1-year gap; both surrounding postal codes present → Case 1 with p = 0.95; k = 6; d = 0 → Rule B	K1A1A1	**K1A*** **	K1A2B2	**DUMMY0**	K1A2B2
3	(empty)	K1A1A1	(empty)	K1A1A1	K1A1A1	Missing postal code in year 1 = 1-year length; no postal code before gap → Case 2c; n/a → Rule B	Missing postal codes in year 3 = 1-year gap; both surrounding postal codes present → Case 1 with p = 0.95; k = 6; d = 0 → Rule A	**DUMMY9**	K1A1A1	**K1A1A1**	K1A1A1	K1A1A1
4	K1A1A1	(empty)	(empty)	(empty)	K1A1A2	Missing postal code in years 2, 3, 4 = 3-year length; both surrounding postal codes present → Case 1 with p = 0.80; k = 5; d = 1 → Rule A for 1st missing postal code; Rule B for 2nd missing postal code; Rule A for 3rd missing postal code	n/a	K1A1A1	**K1A1A***	**DUMMY1**	**K1A1A***	K1A1A2
5	K1A1A1	K1A1A1	K1A1A1	(empty)	(empty)	Missing postal code in years 4, 5 = 2-year length; no postal code after gap → Case 2b; n/a → Rule B used for 2 missing postal codes	n/a	K1A1A1	K1A1A1	K1A1A1	**DUMMY8**	**DUMMY8**
6	(empty)	(empty)	(empty)	(empty)	(empty)	One gap of 5 years → Case 2a; n/a → Rule B used for 5 missing postal codes	n/a	**DUMMY7**	**DUMMY7**	**DUMMY7**	**DUMMY7**	**DUMMY7**
7	K1A1A1	K1A1A1	K1A1A2	K1A1A1	K1A1A1	No gaps → n/a; n/a → No imputation	n/a	K1A1A1	K1A1A1	K1A1A2	K1A1A1	K1A1A1

Notes: The imputed values are in red. Rule A: The missing postal code is imputed based on the postal codes surrounding the gap. Rule B: The missing postal code is given a value unrelated to the surrounding postal codes. k: similarity value—the k characters common to postal codes at both sides of the gap; d: dissimilarity value (equal to 6 minus k); DUMMY0, DUMMY1, DUMMY7, DUMMY8 or DUMMY9: values unrelated to surrounding postal codes given to missing postal codes; n/a: not applicable; p: probability threshold; *: place marker for a missing character in the postal code.

References

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, and S.W. Wang. 2014. “Fifteen-year global time series of satellite-derived fine particulate matter.” *Environmental Science and Technology* 48 (19): 11109–11118.

Peters, P.A., M. Tjepkema, R. Wilkins, P. Fines, D. L. Crouse, P.C.W. Chan, and R. T Burnett, 2013. “Data Resource Profile: 1991 Canadian Census Cohort.” *International Journal of Epidemiology* 42 (5): 1319–1326.

Pinault, L., M. Tjepkema, D.L. Crouse, S. Weichenthal, A. van Donkelaar, R.V. Martin, M. Brauer, H. Chen, and 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.

Statistics Canada. 2014. *Postal Code^{OM} Conversion File Plus (PCCF+) Version 6C, Reference Guide – November 2014 Postal Codes^{OM}*. Statistics Canada Catalogue no. 82-F0086-XDB. Ottawa: Statistics Canada.

van Donkelaar, A., R.V. Martin, J.D. Spurr, and 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): 10482–10491.

Wilkins, R., M. Tjepkema, C. Mustard, and R. Choinière. 2008. “The Canadian census mortality follow-up study, 1991 through 2001.” *Health Reports* 19 (3): 25–43. Statistics Canada Catalogue no. 82-003-XPE.