Misclassification errors from postal code-based geocoding to assign census geography in Nova Scotia, Canada

Mikiko Terashima, PhD,¹ George Kephart, PhD²

ABSTRACT

OBJECTIVES: Postal codes are often the only available geographic identifiers in many sources of health data in Canada. In order to conduct geographic analyses, postal codes are routinely geocoded to census geography to link to ecological data. Despite common use of this method, the extent of geographic misclassification errors is poorly understood. We estimated misclassification errors in the geocoding of postal codes to assign census geography in Nova Scotia, Canada.

METHODS: We examined differences between counts and match rates for postal-code geocoded and actual locations of buildings in Nova Scotia at two census administrative area levels: dissemination areas (DAs) and census subdivisions (CSDs). Actual locations were based on the data collected by the provincial government containing actual latitude/longitude of buildings. Variation in misclassification by rurality, using Statistics Canada's classification, was also assessed.

RESULTS: Outside two urban areas (Halifax Metro and Sydney) which had <10% differences in counts, many DAs had >30% differences. Match rates showed similar patterns, with the vast majority of non-urban DAs having <40% match rates. Even in major urban areas, 10% of DAs had large misclassification errors. Misclassification errors at the CSD level were still too great to estimate counts or rates without further area aggregation.

CONCLUSION: Routine use of postal code geocoding should be replaced with geocoding of location information using additional identifiers such as civic addresses or latitude and longitude. If data holders did this in-house before providing data to researchers, the accuracy and capacity of geographic analysis would be enhanced while protecting confidentiality.

KEY WORDS: Geocoding; postal code; data linkage; small-area analysis; population health

La traduction du résumé se trouve à la fin de l'article.

Can J Public Health 2016;107(4-5):e424–e430
doi: 10.17269/CJPH.107.5459

Geocoding of data for population health research is now a regular occurrence.^{1–3} The word “geocoding” has become a common term in the population health discourse to describe the process of translating location information to an actual geographic location. It may be conducted using different types of location information, such as civic addresses, postal codes and place names. However, postal codes are often the only detailed geographic identifiers attached to research data in Canada, such as administrative health data. Even though finer location information, such as civic addresses, is normally collected for these data, it is not made available to researchers for privacy reasons. With the absence or unavailability of finer location information, postal code-based geocoding has become commonplace in population health research in Canada.

Postal codes have a number of limitations as a geographic identifier for geocoding. First, six-digit postal code area boundaries are poorly defined. Only the first three digits of the postal code (the forward sortation area or FSA) have geographic boundaries for which Statistics Canada provides tabulations. Use of only FSAs as geographic areas limits research because many FSAs are very large. For many research applications, smaller and more socially meaningful geographic areas are required. Second, postal codes do not necessarily indicate place of residence. They may indicate a post office or mailbox – a problem that will

become more common with diminished home delivery. Third, postal codes do not correspond cleanly to standard geographic classifications used by Statistics Canada. Since researchers routinely rely on census or survey data to obtain population denominators or ecological variables, it is important that area-level socio-demographic and health data can be accurately linked between data sources.

To address these problems, Statistics Canada has developed geocoding software and supporting data to impute geographic location in census-defined and electoral areas from six-digit postal codes (PCCF+, or the Postal Code Conversion File Plus). PCCF+ consists of SAS and STATA programs with associated data files.⁴ It allocates postal codes to a variety of different census administrative areas, including dissemination areas (DAs), census tracts and census subdivisions (CSDs). Whenever a postal code could match

Author Affiliations

1. School of Planning, Department of Community Health and Epidemiology, Healthy Populations Institute, Dalhousie University, Halifax, NS
2. Department of Community Health and Epidemiology, Dalhousie University, Halifax, NS

Correspondence: Mikiko Terashima, PhD, School of Planning, Department of Community Health and Epidemiology, Healthy Populations Institute, Dalhousie University, 5410 Spring Garden Rd., Halifax, NS B3H 4R2, Tel: 902-494-3114, E-mail: mikiko.terashima@dal.ca

Conflict of Interest: None to declare.

with more than one area, PCCF+ probabilistically allocates the postal code to one of the possible areas that overlap with the postal code boundary, using a supplemental file of estimated population weights.⁵ PCCF+ also produces an estimated latitude and longitude, which is the centroid of the smallest allocable census area (e.g., a block face in urban areas).

Despite the common use of the PCCF+ in Canadian research for postal code geocoding, the extent of locational misclassification and the potential impacts on research have not been well evaluated.⁶ The effect on research may vary depending on the geographic scope of research and how area data are used. For example, in rural and low density suburban areas, where one postal code can encompass a large geographic area, inaccuracies may be large. In studies of variation in area-level counts, incidence rates or prevalence rates (e.g., raw or age-standardized incidence of injury),⁷ locational misallocations may cancel each other out, limiting bias. On the other hand, in studies assessing area-level effects on health status or outcomes (e.g., regression analysis of association between neighbourhood characteristics and chronic diseases),^{8,9} locational inaccuracies may incorrectly match attributes between data sources. The analysis may be at an area level only,^{10,11} or may be multilevel – involving both area- and individual-level variables.^{9,12,13}

Using highly accurate locational data for all buildings in Nova Scotia, this paper estimated the extent of locational misclassification error, by levels or rurality, resulting from the use of PCCF+ to geocode six-digit postal codes to census geographic areas. We examined the extent of misclassification error employing two measures: 1) the percent difference in counts between actual data and geocoded data in respective areas; and 2) the percentage of geocoded records correctly assigned to actual areas (*match rates*). These measures are most pertinent to the common focus of health research on computation of area-level variation in health and disease occurrences, or the assignment of ecological attributes based on geography.^{14,15}

METHODS

Data and measures

We used a set of building location data maintained by the Government of Nova Scotia, called the Nova Scotia Civic Address File (NSCAF). The NSCAF is a dataset compiled primarily for 911 emergency services, and contains location information for all residential and commercial buildings in the province.¹⁶ All building locations had latitude and longitude and civic address attributes. The latitude and longitude of each building are within each building's footprint, thus even more accurate than the locations assigned using civic addresses (called *address matching*), for instance.

Postal code geocoding of buildings to DAs and CSDs was conducted using Statistics Canada's PCCF+.⁴ We used the most current version available to us at the time of study, PCCF+ 6A, published in 2014.¹⁷ Census areas¹⁸ for actual building data were assigned based on the latitude and longitude of the building location in the NSCAF data using a *spatial join* function within ArcGIS10.2 (Redland, CA, ESRI). Spatial join link point locations (latitude and longitude, in our case) and geographic areas (polygons). DA and CSD area boundary files were available

through the Data Liberation Initiative agreement between Statistics Canada and Dalhousie University.

Many buildings are non-residential or may have mixed uses, and others may contain multiple household units. For these reasons, the counts of buildings are not exactly proportionate to the size of population. The accuracy and completeness of postal codes in the civic address file was very high and unlikely to bias study results. Building records without postal codes ($N = 6,102$ or 1.5%) in the NSCAF data were uncommon. These were excluded, leaving 408,292 records for the analysis. There was no geographic pattern to excluded records, so they are unlikely to affect misclassification estimates. Of the 408,292 records, 139 (0.03%) were not assigned a DA due to either invalid postal codes or new postal codes not recognized in the version of PCCF+. However, they were assigned a CSD using the first three digits of the postal code.

Levels of rurality were determined by Statistics Canada's settlement types assigned to CSDs – regional municipality (RGM), town (T), municipal district (MD), subdivision of county (SC), and Indian reserve (IRI).¹⁸ The level of rurality generally increases in the same order. Two of the three RGMs in Nova Scotia contain large urban centres, Halifax and Sydney. Towns are satellite centres across the province, such as Truro and Yarmouth, and MD and SC are largely rural areas outside of urban centres. IRIs in Nova Scotia may be rural or urban, though grouped together.

Analysis

Several measures to assess the quality of geocoding have been discussed elsewhere.¹⁵ Of those measures, we employed: 1) the percent difference in counts between actual data and geocoded data in respective areas; and 2) the percentage of geocoded records correctly assigned to actual areas (*match rates*). We examined errors using both census DAs and CSDs as area units of analysis. These area units are two of the smallest census geographies for which most census data are publicly available, and are often used in population health research.^{19,20} Then, we examined spatial and statistical distributions of misclassification across urban and non-urban areas of the province. We presented the results in two ways. First, we mapped the percent differences in counts and match rates across areas at both DA and CSD levels to present the spatial distribution. Then we summarized the statistical distribution of these measures by the class of rurality.

RESULTS

There were significant misclassification errors as measured by both percent differences in counts and match rates, with considerable variation in error by area and class of rurality. Differences between actual counts and geocoded counts averaged about 30% at the DA level, and it was not better at the CSD level. About 42% (172,862 of 408,153) of the records were not matched to correct DAs, and 11% (45,639 of 408,292) were not matched to correct CSDs.

According to the 2011 Canadian Census of Population,²¹ the mean population at the DA level in Nova Scotia was 560 (SD = 247, median = 507). There were 1,645 DAs in the province and mean counts of building points from the NSCAF data at the DA level was 250.9 (SD = 129.8, median = 228). While building

counts were smaller than population, they were largely proportionate to population size. The counts at the CSD level varied widely as 2 out of 99 CSDs represent the two most populated urban centres (Halifax and Sydney). Excluding these two (135,835 and 44,647 respectively), the mean count was 2,423.51 (SD = 2,664.14, median = 1,707.5), and building numbers were proportionate to population counts (mean = 4,476.6, SD = 5,093.9, median = 3,357).

Levels of misclassification across areas

Maps in Figures 1a and 1b show 1) the percentage differences in counts (Figure 1a); and 2) match rates across Nova Scotia at the DA level (Figure 1b). In the Halifax Metro and Sydney Areas – the two most urban areas in the province – the difference in counts at the DA level was largely contained within $\pm 10\%$, with a few exceptions in some non-residential DAs (e.g., parks). Match rates were 98% or higher in the two largest centres, as well as in Truro and

Figure 1. (a) Percentage difference in the count²² of geocoded points from the actual counts, by DA. (b) Percentage of geocoded point locations falling into the correct area (match rate), by DA.

*The 18 DAs with no data (due to no building points) are mostly IRIs and not always visible on the map.

Bridgewater – two of the largest towns in Nova Scotia. Outside these urban areas, the magnitude of classification errors varied greatly. DAs with both high and low percent difference are seen throughout the province. A vast extent of areas have match rates of 40% or lower. While it may still be possible to reliably calculate geocoded counts or rates for small areas in larger urban centres, it is simply unfeasible to do so in other parts of the province, or to make a comparison between regions.

Figures 2a and 2b show the distribution of percent differences in counts and match rates at the CSD level. Areas within $\pm 10\%$ difference in counts and high match rates (e.g., 80% and above) cover a majority of area in the province, which is much better than results at the DA level. However, one third of 99 CSDs had more than 30% difference in counts, and just under a half had 60% or lower match rates. Only 24 out of 99 CSDs have match rates of 90% or higher. Low match rates CSDs appear to be those with geographically smaller size – especially towns and Indian reserves. One subdivision of county also had particularly low

match rates, likely because it shares the same postal codes with towns in the same county. To reduce the level of misclassification, further aggregation of CSDs would be necessary in these areas. This further aggregation will impair the ability of analyses to differentiate, for example, between these towns or IRIs from surrounding rural areas.

Levels of misclassification by rurality

Table 1 provides a more detailed distribution of misclassification error in counts and match rates by rurality type at the DA level. One DA was excluded from the summary calculation as it had a geocoded count of 118 points while there was only one actual building. This is likely due to overlapping postal code areas and the population weight assigned to the block face in which the building is located (i.e., the building may be a high-rise apartment with multiple families). The mean percent difference in counts after excluding this outlier was 31.3% (SD = 91.8%, median = 6.7%).

Levels of misclassification differed by level of rurality, but can be substantial regardless of the level of rurality. For example, median percent differences in counts for regional municipalities and towns were small (3% and 7.0% respectively), while DAs falling into municipal districts (19.5%) and subdivisions of county (21.6%) had relatively high median percentage differences. Yet, even in regional municipalities, 10% of DAs had 50% or greater errors in counts and match rates less than 58%. In towns, 25% of DAs had 110% or greater errors in counts and match rates less than 4.3%, while 10% of DAs had greater than 443% errors in counts and match rates less than 1.4%. Such high errors would result in large biases in counts or rates, and bias in the effects of ecological variables on outcomes.

Table 2 presents the CSD-level summary distribution of percent differences in counts and match rates. The extent of misclassification was not necessarily smaller than that of the DA level, depending on the type of rurality. For example, median percentage differences in counts in municipal districts and subdivisions of county were less than half (6.6% and 11.2% respectively) those at the DA level, and the median match rates increased substantially (86.3% and 83.7%). However, the median percent difference in towns jumped from 7.0% to 54.1%, and match rates decreased to 21.5%. This is probably because a large number of high match rate DAs are clustered in a few of these town type CSDs, resulting in the clusters of DAs now counted as one in the summary statistics, and are consequently “overwhelmed” by lower rate CSDs. Both types of misclassification errors in IRIs were largely unchanged (slightly worsened) from those at the DA level, with wider standard deviations than other rurality classes.

DISCUSSION

To our knowledge, this is the first Canadian study to evaluate the quality of postal code geocoding against “gold standard” locational data. The NSCAF’s postal code and address information is maintained regularly by the provincial government for emergency service purposes.¹⁶ These are the most accurate location data available for all buildings in the province, and errors in location would be minimal.

This study found large and pervasive misclassification errors in the geocoding of postal codes to two levels of census geography,

MISCLASSIFICATION ERROR FROM POSTAL CODE-BASED GEOCODING

Table 1. Counts, percent (%) difference and match rate comparison by rurality for dissemination areas (N = 1644)

Rurality class (number of DAs)		Total (N = 1644)	Regional municipalities (RGM) (n = 808)	Towns (T) (n = 213)*	Municipal districts (MD) (n = 223)	Subdivisions of county (SC) (n = 359)	Indian reserves (IRI) (n = 41)
Total counts		408,153	187,644	43,397	71,754	102,951	2407
Counts distribution (actual)	10 th	128.0	98.0	142.0	208.0	181.0	4.0
	Q1	176.0	58.0	164.0	240.0	245.0	8.0
	Median	228.0	132.0	194.0	287.0	256.0	49.0
	Q3	296.0	278.0	224.0	392.0	323.0	173.0
	90 th	404.0	387.0	277.0	485.0	430.0	283.0
	Mean	250.9	232.5	202.6	321.8	288.4	92.6
Counts distribution (geocoded)	(SD)	(129.8)	(210.0)	(60.2)	(112.8)	(134.6)	(103.4)
	10 th	97.0	96.0	41.0	175.0	149.0	19.0
	Q1	160.0	154.0	97.0	231.0	209.0	32.0
	Median	220.0	208.0	159.0	313.0	277.0	96.0
	Q3	317.0	279.0	200.0	412.0	380.0	189.0
	90 th	439.0	393.0	253.0	551.0	509.0	270.0
% Difference in counts	Mean	232.5	232.8	152.2	338.9	311.5	113.2
	(SD)	(132.0)	(134.8)	(75.0)	(154.6)	(170.2)	(89.3)
	10 th	0.5	0.0	0.7	4.3	2.3	4.2
	Q1	2.0	1.0	2.1	10.0	6.7	22.6
	Median	6.7	3.0	7.0	19.5	21.6	46.4
	Q3	22.6	7.3	110.5	35.5	39.0	92.3
% Match rate	90 th	52.3	15.2	442.9	49.5	67.5	158.1
	Mean	31.3	7.3	112.2	29.9	36.2	78.5
	(SD)	(91.8)	(17.1)	(209.9)	(40.7)	(69.7)	(106.7)
	10 th	7.0	58.6	1.4	7.9	5.5	0.0
	Q1	20.9	82.1	4.3	12.8	11.4	0.0
	Median	79.9	95.1	90.8	21.3	22.5	10.5
% Match rate	Q3	97.0	99.2	97.4	40.5	40.2	46.7
	90 th	99.5	100	98.8	77.6	85.7	70.8
	Mean	61.8	85.9	57.5	31.2	31.6	23.6
	(SD)	(37.5)	(21.1)	(45.0)	(26.3)	(28.0)	(30.1)

* One DA with an extreme percent difference in the town category was excluded from the calculation.

Table 2. Counts, percent (%) difference and match rate comparison by rurality for census subdivisions (N = 96)*

Rurality class (number of CSDs)		Total (N = 96)	Regional municipalities (RGM) (n = 3)	Towns (T) (n = 31)	Municipal districts (MD) (n = 12)	Subdivisions of county (SC) (n = 28)	Indian reserves (IRI) (n = 22)*
Total counts		408,268	187,740	43,406	71,764	102,975	2407
Counts distribution (actual)	10 th	63.0	7258.0	464.0	3384.0	1861.0	4.0
	Q1	475.0	7258.0	502.0	3703.5	2336.5	19.0
	Median	1861.0	44,647.0	1100.0	4978.5	3119.5	63.0
	Q3	3825.0	135,835.0	1651.0	7084.5	4265.5	283.0
	90 th	5695.0	135,835.0	3192.0	9531.0	5695.0	347.0
	Mean	4390.0	62,580.0	1400.2	5980.3	3677.7	125.4
Counts distribution (geocoded)	(SD)	(14,694.0)	(66,137.8)	(1069.8)	(3513.2)	(2109.4)	(130.2)
	10 th	70.0	7564.0	94.0	3965.0	1995.0	18.0
	Q1	271.0	7564.0	136.0	4206.0	2432.5	29.0
	Median	2211.0	44,636.0	463.0	5400.0	3346.0	127.0
	Q3	4244.0	135,805.0	1516.0	7298.0	4656.5	271.0
	90 th	7140.0	135,805.0	3246.0	9377.0	5797.0	328.0
% Difference in counts	Mean	4583.7	62,680.0	1028.3	6294.9	3942.0	143.7
	(SD)	(15,037)	(65,995)	(1252.1)	(3673.6)	(2231.7)	(124.7)
	10 th	1.1	0.02	1.1	0.7	2.3	1.7
	Q1	4.2	0.02	1.7	3.1	4.7	14.3
	Median	16.1	0.02	54.1	6.6	11.2	52.6
	Q3	54.5	4.2	75.8	13.1	20.9	71.9
% Match rate	90 th	81.3	4.2	83.8	18.0	31.1	328.0
	Mean	32.0	1.4	45.4	8.8	14.4	143.7
	(SD)	(41.9)	(2.4)	(33.4)	(8.3)	(14.0)	(124.7)
	10 th	3.2	96.5	4.9	80.9	61.1	0.0
	Q1	16.4	96.5	8.5	83.9	75.0	0.0
	Median	77.9	99.0	21.5	86.3	83.7	6.7
% Match rate	Q3	90.4	99.9	94.4	90.3	89.0	58.6
	90 th	97.0	99.9	99.0	93.1	93.7	96.6
	Mean	57.7	98.5	40.0	86.6	79.6	29.4
	(SD)	(37.8)	(1.8)	(38.8)	(4.8)	(15.2)	(38.6)

* Three CSDs (all IRIs) had more than 10 times as many geocoded counts as the actual counts, which skew the distribution of percentage difference in counts. These were excluded from the calculation.

DAs and CSDs. While misclassification errors were small in the majority of DAs and CSDs, especially in regional municipalities and urban centres, a significant percentage of both DAs and CSDs had

large misclassification errors. Twenty-five percent of DAs have percent differences in counts greater than 22% and match rates lower than 20%. Moreover, we found that while these errors are

larger in more rural settings, significant errors occur across all types of geographic areas.

“What is the acceptable rate of accuracy?” is a difficult question to answer.²⁰ One study examined this question in the context of crime rate calculations using Monte Carlo simulation.²³ They found that a match rate of 85% match was the minimum required for a stable calculation.²³ It would be a valuable exercise to conduct a similar analysis using actual disease data. For example, the crime rate study had 5,215 records across 948 small areas, which is 5.5 counts per area on average. Nova Scotia had 6,133 cancer incidents across 1,645 DAs in 2012, 3.7 counts per area on the average, which is not vastly dissimilar to the crime rate study. Hypothetically, if this 85% threshold applies, over one half of 1,645 DAs (898 or 54.5%) will not yield reliable disease rates at the DA level.

Generalization of our findings to population data may be limited by our use of buildings, not persons, for the analyses. All building types were used, including non-residential, and buildings containing multiple households were counted as one. If misclassification error is systematically related to building type, the error rates for areas we reported in this study may be biased. We doubt this is the case, and in any case high density and non-residential buildings are likely concentrated in denser, urban DAs and CSDs where misclassification errors are lower.

It is also important to note that the distribution of spatial misclassification errors in person-level data would be different than reported in this study. We examined misclassification errors at the level of small areas (DAs and CSDs), not persons, and population is concentrated in more densely settled DAs and CSDs that tend to have lower misclassification errors.

We evaluated the accuracy of only the PCCF+ geocoding methodology. Another commonly used postal code geocoding tool in Canada is a commercial product called CanMap Postal Code Suite.²⁴ Similar results for this software are likely, as no tool can overcome the inherent problems of postal codes. Assigned locations based on postal codes will by necessity involve a considerable level of uncertainty.

The implications of spatial misclassification errors for health research using geocoded data will be of greatest concern where areas are the unit of analysis (e.g., studies of small area rate variation).²⁵ Our results show that geocoding using postal codes could substantially bias estimates of numerators or denominators in rate calculations, and may be particularly severe where only the numerator employs geocoding (e.g., rates where denominators are obtained from census data). The implications of misclassification error may be of lower consequence in studies using postal code-linked ecological variables and individuals as the unit of analysis. There are two reasons for this. First, population is more concentrated in DAs or CSDs with lower misclassification error. Second, misclassification will be to adjacent areas that tend to have similar socio-economic and demographic characteristics, especially in dense urban settings. Nevertheless, our results show that misallocation of ecological characteristics to individuals in many areas will be very high, resulting in potential bias in estimated effects of ecological variables. This may be of particular concern in suburban areas, smaller cities and towns where individuals get misallocated to adjacent rural areas.

Should we continue relying on postal code-based geocoding, or should we consider alternative ways to more accurately assign

geographies using more than just postal codes?²⁶ While postal codes remain a viable tool to link data to census geography in a handful of highly populated urban areas in Canada, our findings clearly demonstrate the limitations in their use outside of a few urban areas. Other provinces would likely follow similar patterns.

Rather than continuing to rely on postal code geocoding, efforts should be placed on geocoding using additional locational identifiers, such as civic addresses. A variety of commercial and publicly available tools of geocoding can be used, depending on the types of location information to which researchers have access.²⁷ For example, estimation of location from a civic address with street name – address matching – can determine a point along roads in a map reference file. For this reason, civic addresses typically point to geographic locations closer to actual locations than postal codes.²⁰

For feasibility and confidentiality reasons, we recommend that data holders and not researchers should assume responsibility for improved geocoding. By pre-assigning geographic areas using the best locational information available, and then providing only geocoded area information to researchers, accuracy can be improved while simultaneously reducing confidentiality concerns. This approach will present even smaller confidentiality risks than disclosing six-digit postal code information. It also offers the opportunity for standardizing area geographies across data sources. Nova Scotia is in the process of doing this for administrative health data. A standardized set of small geographic areas akin to neighbourhoods, nested in existing administrative geography across urban and rural regions (similar to the UK's Super Output Areas²⁸), are being developed. Administrative records will be geocoded to this geography using civic address and building data, and will be used by Nova Scotia researchers.

REFERENCES

1. Krieger N, Waterman P, Lemieux K, Zieler S, Hogan JW. Evaluating the accuracy of geocoding in public health research. *Am J Public Health* 2001; 90:1114–16.
2. Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, et al. Geocoding in cancer research: A review. *Am J Prev Med* 2006;30(2S):S16–24. doi: 10.1016/j.amepre.2005.09.011.
3. Auger N, Daniel M, Platt RW, Wu Y, Luo ZC, Choiniere R. Association between perceived security of the neighbourhood and small-for-gestational-age birth. *Paediatr Perinat Epidemiol* 2008;22(5):467–77. doi: 10.1111/j.1365-3016.2008.00959.x.
4. Wilkins R, Peters PA. *PCCF+ Version 5K User's Guide. Automated Geographic Coding Based on the Statistics Canada Postal Code Conversion Files, Including Postal Codes Through May 2011*. Catalogue no. 82F0086-XDB. Ottawa, ON: Health Analysis Division, Statistics Canada, 2012.
5. Peller P. *An Analysis of the Postal Code Conversion File's Use in Research*. Calgary, AB: University of Calgary, 2011; 1–24.
6. Jacques GM. A research agenda: Does geocoding positional error matter in health GIS studies? *Spat Spatio-temporal Epidemiol* 2012;3:7–16. doi: 10.1016/j.sste.2012.02.002.
7. Bell NJ, Schuurman N, Morad Hameed S. A small-area population analysis of socioeconomic status and incidence of severe burn/fire-related injury in British Columbia, Canada. *Burns* 2009;35(8):1133–41. PMID: 19553025. doi: 10.1016/j.burns.2009.04.028.
8. Wang C, Guttmann A, To T, Dick PT. Neighborhood income and health outcomes in infants: How do those with complex chronic conditions fare? *Arch Pediatr Adolesc Med* 2009;163(7):608–15. PMID: 19581543. doi: 10.1001/archpediatrics.2009.36.
9. Zhang X, Onufrak S, Holt JB, Croft JB. A multilevel approach to estimating small area childhood obesity prevalence at the census block-group level. *Prev Chronic Dis* 2013;10:E68. doi: 10.5888/pcd10.120252.
10. Terashima M, Guernsey JR, Andreou P. What type of rural? Assessing the variations in life expectancy at birth at small area-level for a small population province using classes of locally defined settlement types. *BMC Public Health* 2014;14:162. PMID: 24524307. doi: 10.1186/1471-2458-14-162.

11. Pampalon R, Hamel D, Gamache P. Recent changes in the geography of social disparities in premature mortality in Québec. *Soc Sci Med* 2008;67(8):1269–81. PMID: 18639966. doi: 10.1016/j.socscimed.2008.06.010.
12. Matheson FI, Moineddin R, Glazier RH. The weight of place: A multilevel analysis of gender, neighborhood material deprivation, and body mass index among Canadian adults. *Soc Sci Med* 2008;66(3):675–90. PMID: 18036712. doi: 10.1016/j.socscimed.2007.10.008.
13. Terashima M, Rainham DGC, Levy AR. A small-area analysis of inequalities in chronic disease prevalence across urban and non-urban communities in the Province of Nova Scotia, Canada, 2007–2012. *BMJ Open* 2014; 4(e004459):1–10.
14. Armstrong B. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med* 1998; 55(10):651–56. PMID: 9930084. doi: 10.1136/oem.55.10.651.
15. Rhomberg L, Chandalia J, Long J, Goodman J. Measurement error in environmental epidemiology and the shape of exposure-response curves. *Crit Rev Toxicol* 2011;41(8):651–71. PMID: 21823979. doi: 10.3109/10408444.2011.563420.
16. Government of Nova Scotia. *Nova Scotia Civic Address Users Guide*. Halifax, NS: GeoNOVA, 2015.
17. Statistics Canada. *Postal Code Conversion File Plus (PCCF+) Reference Guide*. Catalogue no. 82-E0086-XDB 6A. Ottawa, ON: Statistics Canada, 2014.
18. Statistics Canada. 2011 Census Dictionary. Catalogue no. 98-301-X, 2012. Available at: <http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/index-eng.cfm> (Accessed December 10, 2015).
19. Ross NA, Tremblay S, Graham K. Neighbourhood influences on health in Montreal, Canada. *Soc Sci Med* 2004;28:443–78.
20. Goldberg DW, Jacquez GM. Advances in geocoding for the health sciences. *Spat Spatio-temporal Epidemiol* 2012;3:1–5. doi: 10.1016/j.sste.2012.02.001.
21. Census of Population. Catalogue no. 12-581-X. Available at: <http://www.statcan.gc.ca/pub/12-581-x/2012000/pop-eng.htm> (Accessed November 30, 2015).
22. Iburu S, Fujita J, Yajima H, Kakuda H, Sakamoto M, Matsumura A. The intervention against an outbreak of pulmonary tuberculosis in the dormitory of construction laborers – Connection with approaches from public health, medical treatment, social welfare, and labor management. *Kekkaku* 2001; 76(11):691–98. PMID: 11766360.
23. Ratcliffe JH. Geocoding crime and a first estimate of a minimum acceptable hit rate. *Int J Geogr Inform Sci* 2004;18(1):61–72. doi: 10.1080/13658810310001596076.
24. DMTI Spatial. *Platinum Postal Code Suite v2011.3*. Markham, ON: Multiple Enhanced Postalcodes (MEP), 2011.
25. Kephart G, Asada Y, Atherton F, Burge F, Campbell L-A, Dowling L, et al. *Small Area Variation in Rates of High-Cost Healthcare Use Across Nova Scotia*. Halifax, NS: Maritime SPOR Support Unit, 2016.
26. Fuller D, Shareck M. Canada Post community mailboxes: Implications for health research. *Can J Public Health* 2014;105(6):e453–55.
27. Shah TI, Bell S, Wilson K. Geocoding for public health research: Empirical comparison of two geocoding services applied to Canadian cities. *Can Geogr* 2014;58(4):400–17. doi: 10.1111/cag.12091.
28. Office for National Statistics UK. *Guidance and Methodology, Super Output Areas*. ONS, London, UK.

Received: January 6, 2016
Accepted: July 3, 2016

RÉSUMÉ

OBJECTIFS : Les codes postaux sont souvent les seuls identifiants géographiques disponibles dans de nombreuses sources de données sanitaires au Canada. Afin de procéder à des analyses géographiques, les codes postaux sont habituellement géocodés à la géographie du recensement pour être reliés aux données écologiques. Bien que ce soit une méthode couramment utilisée, on connaît mal l'étendue des erreurs de classification géographique. Nous avons estimé les erreurs de classification dans le géocodage des codes postaux pour fins d'association à la géographie du recensement en Nouvelle-Écosse, au Canada.

MÉTHODE : Nous avons examiné les écarts entre les numérations et les taux d'appariement d'emplacements géocodés selon le code postal et d'emplacements réels de bâtiments en Nouvelle-Écosse à deux niveaux de régions administratives du recensement : les aires de diffusion (AD) et les subdivisions de recensement (SDR). Les emplacements réels ont été déterminés selon les données recueillies par le gouvernement provincial indiquant la latitude et la longitude réelles des bâtiments. Nous avons aussi évalué la variation des erreurs de classification par ruralité à l'aide de la classification de Statistique Canada.

RÉSULTATS : Sauf dans deux agglomérations urbaines (Sydney et la région métropolitaine de Halifax) où il y avait <10 % d'écarts dans les numérations, beaucoup d'AD affichaient des écarts >30 %. Les tendances étaient semblables pour les taux d'appariement : la très grande majorité des AD non urbaines affichaient des taux d'appariement <40 %. Même dans les grandes agglomérations urbaines, 10 % des AD comportaient d'importantes erreurs de classification. Les erreurs de classification à l'échelle des SDR étaient encore trop importantes pour estimer les numérations ou les taux sans un regroupement plus poussé des zones.

CONCLUSION : L'utilisation habituelle du géocodage par code postal devrait être remplacée par le géocodage de l'information de localisation à l'aide d'identifiants supplémentaires, comme les adresses de voirie ou la latitude et la longitude. Si les détenteurs de données faisaient cela à l'interne avant de fournir leurs données aux chercheurs, l'exactitude et la capacité des analyses géographiques seraient rehaussées, et la confidentialité des données serait protégée.

MOTS CLÉS : géocodage; code postal; couplage de données; analyse de données régionales; santé des populations