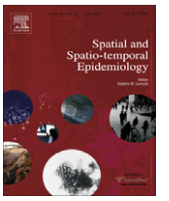
Spatial and Spatio-temporal Epidemiology 3 (2012) 55–67

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste



Quantifying the magnitude of environmental exposure misclassification when using imprecise address proxies in public health research

Martin A. Healy, Jason A. Gilliland *

The University of Western Ontario, 1151 Richmond St. N, London, Ontario, Canada N6A 5C2

ARTICLE INFO

Article history:
Available online 11 February 2012

Keywords:
Geographic information systems
Geocoding
Accessibility
Environmental health
Public health

ABSTRACT

In spatial epidemiologic and public health research it is common to use spatially aggregated units such as centroids of postal/zip codes, census tracts, dissemination areas, blocks or block groups as proxies for sample unit locations. Few studies, however, address the potential problems associated with using these units as address proxies. The purpose of this study is to quantify the magnitude of distance errors and accessibility misclassification that result from using several commonly-used address proxies in public health research. The impact of these positional discrepancies for spatial epidemiology is illustrated by examining misclassification of accessibility to several health-related facilities, including hospitals, public recreation spaces, schools, grocery stores, and junk food retailers throughout the City of London and Middlesex County, Ontario, Canada. Positional errors are quantified by multiple neighborhood types, revealing that address proxies are most problematic when used to represent residential locations in small towns and rural areas compared to suburban and urban areas. Findings indicate that the shorter the threshold distance used to measure accessibility between subject population and health-related facility, the greater the proportion of misclassified addresses. Using address proxies based on large aggregated units such as centroids of census tracts or dissemination areas can result in very large positional discrepancies (median errors up to 343 and 2088 m in urban and rural areas, respectively), and therefore should be avoided in spatial epidemiologic research. Even smaller, commonly-used, proxies for residential address such as postal code centroids can have large positional discrepancies (median errors up to 109 and 1363 m in urban and rural areas, respectively), and are prone to misrepresenting accessibility in small towns and rural Canada; therefore, postal codes should only be used with caution in spatial epidemiologic research.

1. Introduction

Recent advances in the analytical capacity of desktop geographic information system (GIS) software, combined with the increasing availability of spatially-referenced health and environmental data in digital format, have created new opportunities for making breakthroughs in spatial epidemiology (Zandbergen, 2008). As digital mapping is an abstraction of reality, the spatial data used for visualizing and analyzing geographic phenomena will always be inaccurate to some degree. Such inaccuracies can be compounded when spatially aggregated units are used as locational proxies for mapping and analyzing spatial relationships, rather than more precise geographic locations. In environmental and public health research, it is common to use proxies for sample unit locations, such as centroids of postal/zip codes, census tracts, dissemination areas, blocks, or lots; however, it is very uncommon for

* Corresponding author. Tel.: +1 (519) 661 2111x81239; fax: +1 (519) 661 3750.
E-mail address: jgillila@uwo.ca (J.A. Gilliland).

1877-5845/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sste.2012.02.006

studies to address, or even mention, the potential problems ensuing from the positional discrepancies associated with using imprecise address proxies. It is the responsibility of the researcher to identify, quantify, interpret, and attempt to reduce any errors associated with using particular spatial data and locational proxies, so that they do not interfere with any conclusions and recommendations to be made from the findings (Fotheringham, 1989; Anselin, 2006).

Researchers in spatial epidemiology have long been concerned about the absolute or relative spatial accuracy of the address points used to map sample populations or phenomena within a GIS (Goldberg, 2008). Numerous researchers have examined the 'positional errors' which occur when the address from a database is located on a digital map, but the point is not located at the true position of the address (Cayo and Talbot, 2003; Ward et al., 2005; Schootman et al., 2007; Strickland et al., 2007; Zandbergen and Green, 2007; Jacquez and Rommel, 2009). In many previous studies, positional errors are reported as Euclidian distance errors, or errors in the X and Y dimension. While much has been said about positional errors, much less has been said about how study results might be affected when researchers use spatially aggregated units (which themselves might be positionally accurate) as address proxies. Very few studies measure and compare the positional discrepancies between address proxies and the exact address they are used to represent (Bow et al., 2004).

A major area of investigation in the fields of spatial epidemiology, health geography, and public health attempts to assess the levels of accessibility or 'exposure' of subject populations to elements in their local environments that are believed to be health-promoting or health-damaging, and are related to certain health-related behaviors or outcomes. Accessibility is typically measured in relation to the distance between subject populations and selected environmental features, and is often operationalized as a binary variable (i.e., accessible/inaccessible, exposed/not exposed) or a density variable (i.e., number of sites within, volume of contaminant within) in relation to an areal unit or 'buffer' of a certain threshold distance (radius) around the subject's address. There is much variability, but unfortunately not much debate, regarding the particular threshold distances to be used in accessibility studies; however, most authors do attempt to justify their choice of threshold distances based on human behavior (e.g. 'walking distance') or perhaps some characteristic of contaminant source (e.g. 150 m from roadway). The chosen accessibility thresholds also typically vary by study population (e.g. children vs. adults), setting (e.g. urban vs. rural), and by health-related outcome (e.g. physical activity vs. asthma). In their study of the environmental influences on whether or not a child will walk or bike to school, for example, Larsen and colleagues (2009) justify the choice of a 1600 m neighborhood buffer based on the local school board cut-off distance for providing school bus service (see also Schlossberg et al., 2006; Muller et al., 2008; Brownson et al., 2009; Panter et al., 2009). Studies which have focussed on access to neighborhood resources such as public parks and recreation spaces have utilized a variety of threshold distances, typically between 400 and 1600 m (compare Lee et al., 2007; Bjork et al., 2008; Tucker et al., 2008; Maroko et al., 2009); however, we submit a threshold distance of 500 m is ideal, as it represents a short 5–7 min walk, therefore easily accessible for populations of all ages (see Tucker et al., 2008; Sarmiento et al., 2010; Wolch et al., 2010). The 5–7 min walk zone, as represented by the 500 m buffer around a home or public school, is also a common distance used in studies exploring the relationship between access to junk food and obesity (see Austin et al., 2005; Morland and Evenson, 2009; Gilliland, 2010). Studies of 'food deserts' (disadvantaged areas with poor access to retailers of healthy and affordable food) and the potential impact of poor access to grocery stores on dietary habits and obesity have tended to focus on longer distances (800 m or greater), and vary according to urban vs. rural setting (see Wang et al., 2007; Larsen and Gilliland, 2008; Pearce et al., 2008; Sharkey, 2009; Sadler et al., 2011). For the purpose of this analysis, we focus on 1000 m, or the 10–15 min walk zone around a grocery store, as has been identified in previous studies of food deserts in Canadian cities (Apparicio et al., 2007; Larsen and Gilliland, 2008). Explorations of how distance from a patient's home to emergency services available at hospitals is associated with increased risk of mortality are more likely to use much larger threshold distances than standard 'walk zones' (e.g. greater than 5 km) (see Jones et al., 1997; Cudnick et al., 2010; Nicholl et al., 2007; Acharya et al., 2011). Nicholl and colleagues (2007), for example, discovered that a 10 km increase in straight-line distance to hospital is associated with a 1% increase in mortality. As hospitals tend to be a regional, rather than a neighborhood facility, we will use the threshold distance of 10 km for our analyses.

Rushton and colleagues (2006) have argued that when short distances between subject population and environmental features are associated with health effects in epidemiologic studies, the geocoding result must have a positional accuracy that is sufficient to resolve whether such effects are truly present. The purpose of this study is to quantify the magnitude of the positional discrepancies in terms of distance errors and accessibility misclassification that result from using several commonly-used address proxies in public health research. Positional errors have been shown to vary greatly by setting (Bonner et al., 2003; Cayo and Talbot, 2003; Ward et al., 2005); therefore, we quantify errors by multiple neighborhood types: urban, suburban, small town, and rural. We also attempt to ascribe 'meaning' to these errors for spatial epidemiologic studies by examining errors in distance and accessibility misclassification with respect to several health-related features, including hospitals, public recreation facilities, schools, grocery stores, and junk food retailers.

2. Methods

2.1. Study area and data

The City of London (population 350,200) and Middlesex County (population 69,024) in Southwestern Ontario, Canada are ideal study areas for examining the geocoding errors in accessibility studies as they encompass a mix of urban, suburban, small town, and rural agricultural areas

(Statistics Canada, 2011) (see Fig. 1). The study area was categorized into four neighborhood types as follows: (1) *urban* areas correspond to neighborhoods in the City of London built primarily before World War II; (2) *suburban* neighborhoods are areas built following WWII that fall within London's contemporary urban growth boundary; (3) *small towns* are settlements outside London within Middlesex County, these settlements have fewer than 20,000 inhabitants; and (4) *rural* areas are defined as all areas of Middlesex County not identified as small town, as well as areas within the city limits of London which are outside its urban growth boundary. All of the areas combine for a total of 104,025 residential addresses, as well as 94 census tracts, 665 dissemination areas and population weighted dissemination areas, 1410 dissemination blocks, 14,256 postal codes, and 19,365 street segment center points. The spatial relationship between geographically aggregated units and a sample dwelling centroid are illustrated in Fig. 2. The dwelling centroid is located within hierarchical spatial structure starting with the census tract, moving down to dissemination area, and then to the dissemination block and finally the individual parcel of land or lot. The dwelling unit is also located within a postal code region, and on a street segment. Each of these larger geographic units can be operationalized as point locations according to their centroids, as seen in Fig. 2.

Digital spatial layers to be used as our address proxies were prepared in ArcMap-ArcInfo10.0 (ESRI Inc., 2011). The census tract, dissemination area, and dissemination block boundary files, supplied by Statistics Canada (2006), were converted to centroids using the 'Feature to Point' tool. These three spatially aggregated units are commonly used in geographic analyses of population data in Canada; each having their own tradeoffs for researchers based on the size of the aggregated unit vs. the richness of data available. Dissemination blocks are the smallest of the three geographic units in terms of area; therefore their centroids provide a more spatially accurate proxy for exact address. However, most Canadian census data, except population and dwelling counts, is suppressed at this level, and for this reason, the utility of dissemination blocks in studies of accessibility among population subgroups is more limited. Dissemination areas are made up of a small group of dissemination blocks. They are commonly-used in population health studies as they are the smallest aggregated geographic unit available for which Statistics Canada releases a number of key demographic variables (e.g. median household income, population by age, population by ethnicity); nevertheless, a considerable amount of data suppression still occurs at this scale. While census tracts are the most commonly-used proxy for 'neighborhoods' in sociological, geographical, and population health research in Canada, and they offer the most comprehensive census data for spatial epidemiologic analyses, they are also the largest geographic unit examined in this study. For this reason, they are hypothesized to result in the greatest positional discrepancy when used as address proxies. Additionally, census tracts are only available in metropolitan areas and therefore do not cover most rural areas. The weighted dissemination areas centroids were created using the 'Median Center' tool by leveraging the population distribution data stored within dissemination block centroids which were nested within the dissemination areas. The weighted dissemination areas centroid has been used in previous research (e.g. Apparicio et al., 2008; Henry and Boscoe, 2008) and was included in this study as a more representative measure for the probable location of population within the area. It is therefore expected to produce a closer approximation for an address proxy than the dissemination area centroid. The postal code boundaries and points were drawn from the Platinum Postal Code Suite (DMTI Spatial Inc., 2009). The typical postal code in a Canadian city is a much smaller geographic unit than the typical US zip code, and is commonly used as a proxy for residential address by Canadian researchers when full civic address is unavailable, or suppressed to maintain subject privacy (e.g. Larsen et al., 2009). The street segment centers were created using the tool 'Feature

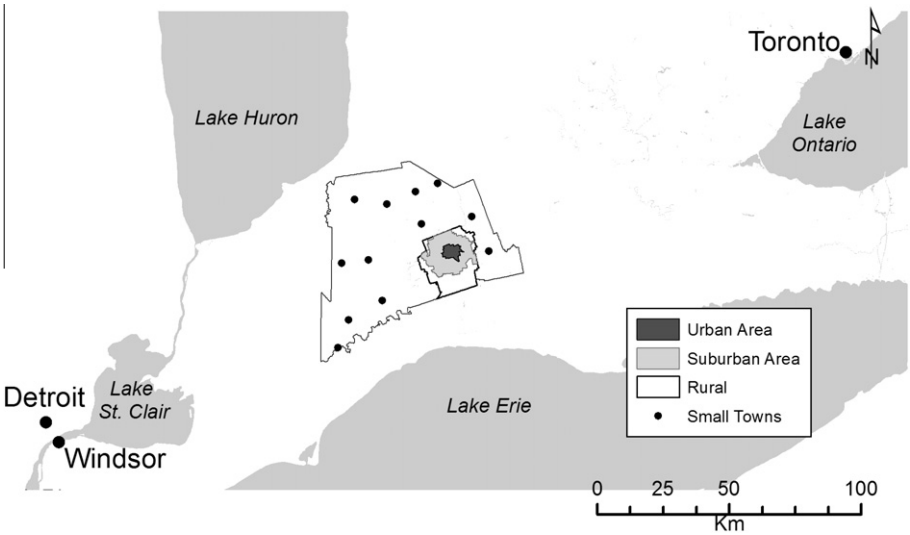


Fig. 1. Study area: London and Middlesex County, Ontario.

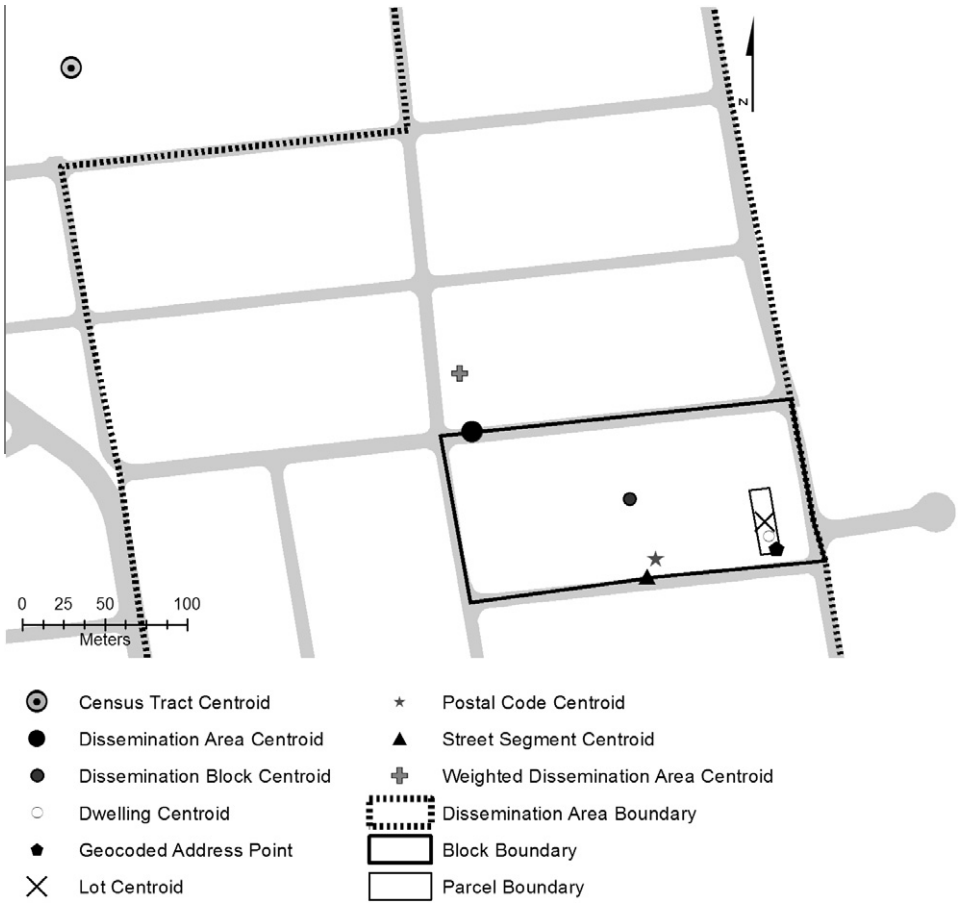


Fig. 2. Spatial relationships between various geographic aggregation levels and their corresponding centroid within a census tract.

Vertices to Points' with the CanMapstreet files (DMTI Spatial Inc., 2009). The geometric center of every street segment was generated as an aggregate address proxy for all the dwellings on that segment. The average street length for rural neighborhoods was 711 m, 187 m for small towns, 142 m for suburban neighborhoods, and only 127 m for urban neighborhoods. All 147,000 addresses points in the study area were supplied by the City and County for every parcel of land, dwelling, business, and institution (City of London, 2010; Middlesex County, 2011). A total of 104,025 address points were identified as residential, and each point was located within the centroid of the dwelling polygons provided by the City and County. A tabular list of each of the residential addresses was generated and these addresses were used to geocode against the CanMap street files (2009) using the 'US Address – Dual Ranges' address locator, thus generating interpolated address points with the default 10 m offset from the street center line. These interpolated addresses, referred to as 'geocoded points' in this paper, are undeniably the most commonly-used address proxies when full address information is available to the researcher. While most researchers use such geocoded points without question, we argue that even these address proxies could have positional discrepancies which might cause accessibility misclassification and therefore they must also be subjected to further scrutiny. Dwelling centroids are the 'gold standard' of address proxies in this study, to which all other address proxies will be measured. We submit that this is the best choice, as all journeys from the home begin somewhere within the home. In this paper, the issues of address validity and match rates for dwelling and lot centroid are controlled for, in that every one of the 104,025 residential addresses were matched at 100%. To calculate accessibility measures, the centroids for dwelling centroids and all the address proxies (except those located on the street segment or a fixed distance from the street segment) were linked with a connecting lateral line from the proxy address point to the nearest corresponding street segment using a custom algorithm. These lateral lines were included in the network distances reported in the study. The street segment center points already located on the street centerline did not require a lateral line to connect them to the network, while the geocoded points were all standardized to be 10 m from the street centerline and thus the 10 meters were added to the individual distances post process.

GIS layers including the locations of all 6 hospitals, 138 elementary schools, and 512 public recreation spaces within the study area were provided by the geomatics divisions of the City and County (City of London, 2010; Middlesex County, 2011). Addresses for the 52 grocery stores and 1213 junk food retailers (including fast food restaurants

and convenience stores) in the study area were provided by the Middlesex-London Health Unit (MLHU, 2010) and geocoded using the master address files provided by the City and County. All data was verified and corrected using orthorectified air photos of London and Middlesex (15 and 30 cm resolution, respectively) (City of London, 2010; Middlesex County, 2011). For built structures, the centroid of the building polygon was used as the address 'gold standard'; however, for recreational places without a defined built structure, such as parks, the access points were manually created using the air photos. The City, County, DMTI Spatial Inc., and Statistics Canada publish no metric regarding the absolute or relative spatial accuracy of their datasets. In this study, the City and County spatial data were accepted as the most spatially accurate of all the data sources. The City and County spatial data were used to create the building centroids for facilities, dwellings, and the centroid for dwelling lots. Spatial features found in the Statistics Canada and DMTI Spatial Inc. data are within 15 m of the same corresponding features in the City and County data for most of the study area. The Statistics Canada and DMTI Spatial Inc. data were used to generate the census tract, dissemination area, weighted dissemination area, dissemination block, postal code centroids, the street segment center, and the geocoded point address proxies, and to generate the shortest path network routes and polygons.

2.2. GIS methods

Shortest path routes (by distance) along the street network from the address proxies to the health-related destination facilities were created using the ArcMap 10.0 Network Analyst 'Closest Facility' function (ESRI Inc., 2011). Starting from each dwelling centroid a network route was created to the nearest health-related facility (i.e., the nearest hospital, school, grocery store, junk food outlet, and public recreation facility). This procedure was repeated for every type of health facility until all 104,024 dwelling centroids were assigned a separate shortest path route to one of each of the facility types. The process was then repeated for each of the eight address proxies. The distance measures were stratified into rural, small town, suburban, and urban neighborhood types and exported from ArcMap 10.0 for analysis in Excel 2010 (Microsoft, 2011) and PASW 18 (IBM, 2011). A recent study of accessibility to multiple food retailer types in rural Middlesex County illustrated how accessibility can be misclassified if facilities outside the county boundary are not considered in distance calculations (Sadler et al., 2011). Sadler and colleagues (2011) demonstrated that when facilities in neighboring counties were included in the spatial analyses, distance to the nearest grocery store decreased for nearly one-third of households, and distance to nearest fast food outlet decreased for over one-half of households. The edge effect was taken into account in the present study by compiling the datasets for selected health-related facilities in neighboring counties (within 10 km from the border of Middlesex County) and then including these facilities in the distance calculations.

2.3. Misclassified address proxies

When spatial aggregations of the subject populations or geographic features are used as proxies in a study of accessibility, the researcher risks misrepresenting the accessibility metric used in that study. Fig. 3 illustrates several potential problems of misclassification and miscounting of grocery stores by identifying three accessibility areas; the census tract boundary; a1000 m network service area buffer originating from the centroid of that same census tract; and a 1000 m network service area buffer originating from a dwelling centroid from within the same census tract. The figure shows that the census tract boundary and the 1000 m network service area buffer around the census tract centroid does not contain a grocery store, and thus would be coded as inaccessible; however, the dwelling centroid buffer does 'contain' at least one grocery store and would be coded as accessible. Fig. 3 also illustrates that the count and density metrics will be affected by the positional discrepancy of using imprecise address proxies. We see that the census tract boundary and the buffer around the census tract centroid do not contain any grocery stores, while the dwelling centroid buffer contains two grocery stores. A further look at the Fig. 3 reveals that the distance between the census tract centroid and the dwelling centroid is biased in the direction of the positional discrepancy. In this example, if the census tract centroid was used as the address proxy, the researcher would have coded all sample unit locations within the census tract as not having a grocery store within 1000 m, when in fact, there are two grocery stores within 1000 m for some of the sample units. Moreover, the researcher would have over-estimated the distance to the closest grocery store for numerous dwelling units, such as the one in our example.

Following some commonly-used distances found in previous health-related studies of accessibility (as noted above), the thresholds distances used in this study were: 500 m for junk food and public recreation spaces, 1000 m for grocery stores, 1600 m for schools, and 10 km for hospitals. Shortest path route buffers had been created for each address proxy and each address proxy point was binary encoded, either the address proxy was inside the threshold (coded as 1) or outside the threshold (coded as 0). We then matched the binary variable to every dwelling centroid from every corresponding address proxy, and then reported the percentages of improperly coded addresses.

2.4. Statistical methods

The distance discrepancies were generated by taking the shortest path distance from a dwelling centroid to a health-related facility and then subtracting the corresponding shortest distance from each corresponding address proxy to that same health facility type. The Phi correlation coefficient was generated in PASW 18 (IBM, 2011) and was used to measure the association between the binary threshold values (i.e., accessible/inaccessible) between the dwelling centroid threshold value (0,1) to each of its corresponding address proxy threshold values (0,1). Phi will return an association coefficient between −1 and +1. A positive value of +1 occurs when all the dwelling

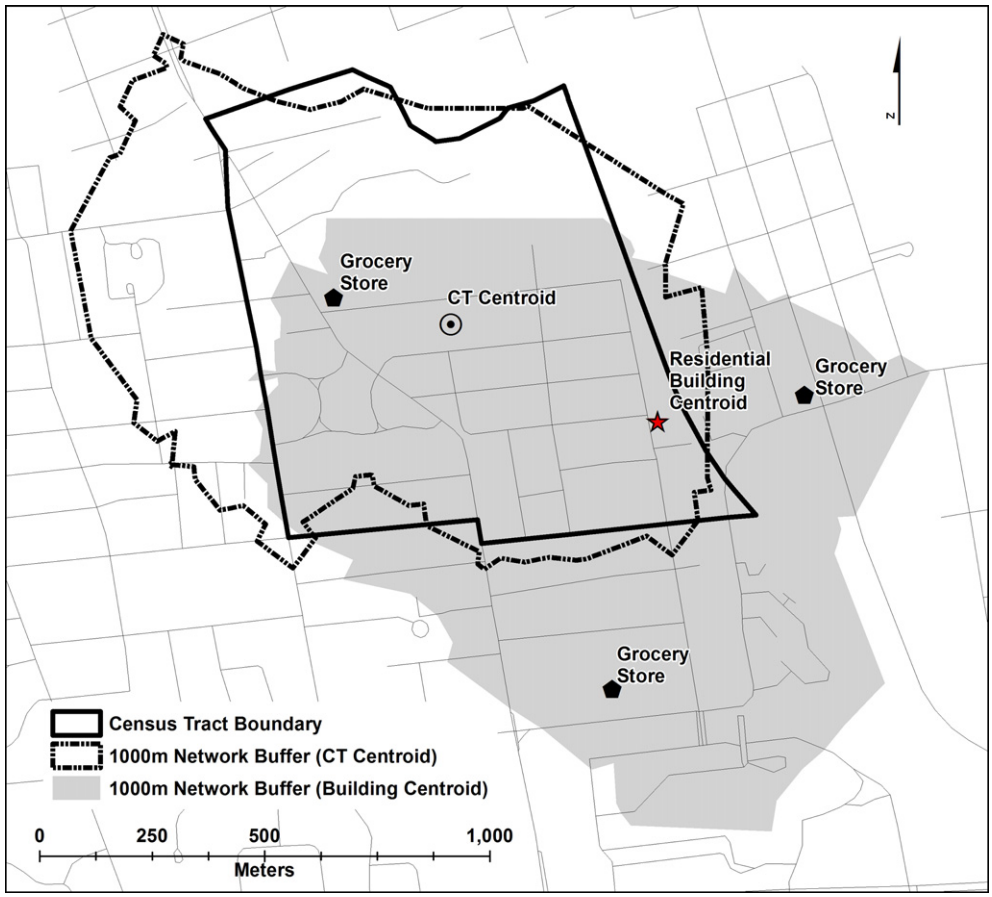


Fig. 3. Illustration of threshold distance miscoding errors.

threshold values and all the address proxy threshold values are in concordance with one another. Conversely, if there is total discordance between all the dwelling threshold values and all the address proxy threshold the Phi coefficient will be −1. If a number of dwelling centroid threshold values differ from those of the corresponding address proxy, the coefficient will begin to move toward 0, thus suggesting a weaker association in terms of accessibility encoding for that address proxy. The significantly positive associations (sig. < 0.01) are between 0.7 and 1.0.

3. Results

3.1. Magnitude of positional discrepancies

In almost every case, urban neighborhoods show the smallest median distance error for all address proxies, followed successively by suburban, small town, and rural areas (see Table 1). As expected, lot centroids were the most accurate proxy for precise residential dwelling location that we examined in relation to nearest distance to health related facilities, with the median positional discrepancy (50th percentile) between lot centroids and dwelling centroids equal to 6–9 m for locations in urban and suburban neighborhoods, 25–43 m for locations in small towns, and 43–50 m for locations in rural areas. The second most accurate proxy for residential location was the geocoded point, with median positional discrepancies between geocoded points and dwelling centroids between 38 and 84 m for residential locations in urban neighborhoods, 37–80 m for locations in suburban neighborhoods, 34–82 m for small town locations, and 77–100 m in rural locations. The third most accurate address proxy we examined was the street segment centroid, with median positional discrepancies in relation to dwelling centroids between 52 and 102 m for residential locations in urban neighborhoods, 75–106 m for locations in suburban neighborhoods, 52–100 m for small town locations, and 173–197 m in rural locations. In urban and suburban areas, the positional discrepancies between postal code centroids and dwelling centroids are very similar to the positional discrepancies between street segment centroids and dwelling centroids; however, the positional discrepancies are drastically worse when using postal codes in small towns (median distance errors between 373 and 1177 m) and rural areas (distance errors between 762 and 1363 m). In rural areas and small towns, the positional errors are always greater when using postal code centroids as address proxies compared to centroids of dissemination blocks, weighted dissemination areas, and dissemination areas. Conversely, postal codes show smaller positional errors than these same address proxies in urban and suburban areas. Census tract centroids are always the

Table 1
Median positional discrepancy (meters) by facility type and neighborhood type.

Neighborhood type	Rural (m)	Small town (m)	Suburban (m)	Urban (m)
Junk food				
Lot centroids	49	29	9	8
Geocoded point	85	48	51	38
Street segment center	175	65	75	52
Postal code	762	373	78	54
Dissemination block	680	147	127	78
Weighted dissemination area	897	279	168	100
Dissemination area	1054	509	176	113
Census tract	930	1414	243	160
Public recreation places				
Lot centroids	43	43	8	8
Geocoded point	77	34	75	84
Street segment center	185	52	106	102
Postal code	896	1177	114	109
Dissemination block	677	156	176	145
Weighted dissemination area	988	296	228	185
Dissemination area	1070	599	241	207
Census tract	1347	1723	352	247
Grocery stores				
Lot centroids	43	25	6	9
Geocoded point	100	82	80	59
Street segment center	197	95	100	76
Postal code	1196	494	98	79
Dissemination block	810	169	141	112
Weighted dissemination area	1193	335	198	145
Dissemination area	1263	559	201	158
Census tract	1704	1870	373	343
Schools				
Lot centroids	50	32	6	9
Geocoded point	94	51	60	55
Street segment center	173	66	80	66
Postal code	913	711	82	68
Dissemination block	665	148	133	101
Weighted dissemination area	1017	361	187	132
Dissemination area	1140	573	194	140
Census tract	1268	1679	363	251
Hospitals				
Lot centroids	46	27	5	8
Geocoded point	85	65	37	75
Street segment center	187	100	67	93
Postal code	1363	537	78	101
Dissemination block	769	349	176	160
Weighted dissemination area	1350	415	203	166
Dissemination area	1255	538	204	171
Census tract	2088	2166	445	343

address proxy with the largest positional error for all neighborhoods and facility types, with median positional discrepancies ranging from the lowest distance error of 160 m (when calculating distance to junk food locations in urban areas) to a high of 2088 m (when calculating distance to hospital in rural areas). Tables A1–A5 (in the Appendix A) provide additional information on the positional discrepancies (including mean distance errors, as well as errors at 75th, 90th, 95th, and 99th percentiles) between the address proxies and the dwelling centroids they are meant to represent. The general pattern observable for the median (i.e., 50th percentile) positional discrepancies (reported in Table 1) tends to be similar in relative terms, but much less dramatic in terms of absolute distance errors, compared to the mean positional discrepancies, as well as the 75th, 90th, 95th, and 99th percentile of discrepancies.

3.2. Positional discrepancy by facility type

The positional discrepancies between the address proxy locations and the dwelling centroids they are to represent not only vary considerably by neighborhood type, but they also vary by health facility type. When lot centroids are used as address proxies, there is a very small variability between distance errors for all facility types, regardless of neighborhood type (rural = ±7 m; urban = ±1 m) (see Table 1). Of the 32 unique combinations of address proxies, neighborhood types, and facility types it is the junk food outlets ($N = 1213$) that have the minimum median distance errors 68.8% of the time (22/32), while public recreation facilities ($N = 512$), singularly, account for almost 50% (15/32) of the facilities with maximum median distance errors. The junk food outlets have minimum median distance errors for all the address proxies in the urban neighborhood type. Junk food outlets, also, account for all the minimum median distance errors in suburban and small town neighborhood types for postal codes, dissemination block, weighted dissemination area, dissemination area, and census tract proxies. For rural neighborhoods, the minimum median distance errors for junk food outlets are found when the postal code, weighted dissemination area, and census tract proxies are used. For the most part, public recreation facilities ($N = 512$) display larger median positional discrepancies than all other health related facilities in urban and suburban areas, while hospitals ($N = 6$) and grocery stores ($N = 52$) show the greatest positional discrepancies compared to the other health related facilities in rural and small towns. The postal code median distance error of 1177 m for small town and public recreation facilities is a larger error than rural neighborhood types and public recreation facilities (896).

3.3. How positional discrepancy impacts accessibility measures

In addition to reporting the positional discrepancy errors it is instructive to look at how much of an effect these errors have on the classification of the population aggregated in each of the address proxies. In some health related accessibility studies continuous variables are used to measure the proximity of health related facilities to an address proxy. Some studies use binary variables to identify whether or not a health related facility exists within a set threshold distance (or buffer radius) around a proxy (Talen, 2003; Apparicio et al., 2008); still more studies use density and counts, however, as indicated in Fig. 3, this approach can also lead to serious misclassification errors. Table 2 considers the impact of positional discrepancy on accessibility, by reporting the percentage of cases that are incorrectly classified as accessible or not, by address proxy, neighborhood type and health related facility type. The general trend is that the smaller the distance threshold, the greater the percentage of addresses misclassified; also, the larger the geographic area of the unit of aggregation, the greater the percentage of addresses which are

Table 2
Accessibility thresholds: percentage of misclassified observations by address proxy.

Address proxy	Neighborhood type	Junk food (500 m)	Recreation places (500 m)	Grocery (1 km)	Schools (1.6 km)	Hospitals (10 km)
Census tracts (N = 94)	Rural* (n = 17)	13.5	8.0	4.7	18.3	26.4
	Small town (n = 3)	36.7	33.7	21.0	35.7	10.2
	Suburban (n = 54)	31.2	47.4	16.8	15.9	5.1+
	Urban (n = 20)	16.9	49.5	37.1	0.1+	0.0+
DA(N-665)	Rural (n = 125)	7.6	3.7	3.8	11.9	8.6
	Small town (n = 43)	35.4	37.1	22.8	29.3	2.7
	Suburban (n = 367)	23.9	28.2	11.4	7.6	0.7+
	Urban (n = 130)	15.5	33.5	15.3+	0.1+	0.0+
Weighted DA (N = 665)	Rural (n = 110)	9.6	4.7	3.9	11.9	8.5
	Small town (n = 53)	31.5	33.5	15.2	19.2	1.2+
	Suburban (n = 372)	23.0	29.2	11.5	6.7	0.7+
	Urban (n = 130)	10.7	29.7	15.5+	0.1+	0.0+
DB (N = 4210)	Rural (n = 1499)	6.9	2.9	2.5	8.5+	5.2+
	Small town (n = 593)	18.2	22.2	11.2+	15.3+	1.0+
	Suburban (n = 1409)	18.4	25.6	9.1	5.9+	0.9+
	Urban (n = 709)	12.0	24.5	13.1+	1.1+	0.0+
Postal code (N = 14,256)	Rural (n = 2539)	9.2	6.8	3.0+	8.1+	6.9+
	Small town (n = 1003)	29.9	33.2	27.8	37.0	3.5+
	Suburban (n = 7792)	11.3+	21.0	6.4+	2.4+	0.3+
	Urban (n = 2922)	6.5	22.8	10.5+	0.1+	0.0+
Street segment (N = 19,365)	Rural (n = 6310)	4.3+	2.3+	1.0+	3.6+	1.2+
	Small town (n = 2227)	9.0+	8.7+	4.3+	2.7+	0.6+
	Suburban (n = 8364)	12.4+	21.6	6.8+	2.3+	0.3+
	Urban (n = 2464)	6.2	23.1	10.1+	0.1+	0.0+
Geocoded (N = 104,024)	Rural (n = 16,686)	2.9+	1.9+	1.1+	2.5+	0.5+
	Small town (n = 14,139)	7.1+	6.7+	4.0+	2.2+	0.4+
	Suburban (n = 54,579)	8.9+	18.3	5.3+	1.5+	0.2+
	Urban (n = 18,620)	5.6+	21.1	9.4+	0.1+	0.0+
Lot (N = 104,024)	Rural (n = 16,686)	0.8+	0.4+	0.2+	0.6+	0.5+
	Small town (n = 14,139)	2.0+	1.8+	0.8+	0.6+	0.1+
	Suburban (n = 54,579)	1.7+	1.5+	0.6+	0.4+	0.1+
	Urban (n = 18,620)	1.5+	1.7+	1.3+	0.0+	0.0+

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of address proxies; n – number of address proxies by neighborhood type.
* Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.
+ Phi coefficient strong positive association (+0.7 to +1.0) sig. < 0.01.

misclassified. For example, using the centroid of a large aggregated unit such as a census tract as a proxy for precise residential address when calculating whether or not a park is located within 500 m from residential addresses in urban neighborhoods will result in nearly half (49.5) of all observations being misclassified. On the other hand, using a large threshold distance of 10 km to determine accessibility to a hospital results in no misclassification in urban areas, no matter what the address proxy used (as the threshold practically covers the entire urban area). The Phi coefficient shows a positive association between each of the dwelling centroids and each and every corresponding address proxy of the coding threshold (inside/outside) across all the health related facility thresholds, except for one. There is a weak negative ($\Phi = -0.6$, $p < 0.01$) association for the urban census tract proxy coding thresholds for public recreation facilities. For example, census tract centroids coded as 'outside' (those that do not have a public recreation facility within 500 m) will have many corresponding dwelling centroids coded as 'inside' (those that do have a public recreation facility within 500 m) resulting in this negative association. There is a strong positive association between dwelling centroid and lot centroid for threshold distances of 1 km to grocery stores. If a suburban dwelling centroid is coded as being within 1 km from a grocery store (code = 1) there is a strong probability ($\Phi = 0.996$, $p < 0.01$) that the corresponding lot centroid will also be within 1 km of a grocery store and coded in the same way. Conversely, if a dwelling centroid is coded as being farther away than 1 km from a grocery store (code = 0) then there is the same probability ($\Phi = 0.996$, $p < 0.01$) that the corresponding lot centroid will also be coded in the same way. The range of Phi values for dwellings and corresponding census tracts, dissemination areas, and weighted dissemination area proxies for junk food and recreation places (500 m thresholds) are weakly associated ($-0.6 < \Phi < 0.47$, $p < 0.01$). The fewest misclassification errors and strongest associations for the 500 m thresholds exist for lot centroids ($\Phi > 0.93$, $p < 0.01$) followed by geocoded points ($0.6 < \Phi < 0.87$, $p < 0.01$). Postal code

centroids showed very high errors in coding for small town (29.9%) and weak association (rural $\Phi = 0.26$, small town $\Phi = 0.29$, suburban $\Phi = 0.59$, and urban $\Phi = 0.58$, $p < 0.01$).

4. Discussion

It is common in public health research to use spatially aggregated units as address proxies for the locations of subjects and facilities when more precise address information is unavailable. It is rare, however, for public health researchers to examine, or even mention, the potential distance and misclassification errors resulting from the positional discrepancies between the locations of imprecise address proxies and precise subject locations. It is inappropriate for researchers to ignore these inaccuracies or to merely accept them as an inevitable component of doing spatial research. It is important to identify and quantify any spatial errors so that we can critically examine research findings and properly advise those to whom policy recommendations are made regarding the potential correlations between subject populations and environmental exposures.

One of the contributions of our study is to quantitatively describe the magnitude of distance errors that result when several of the most commonly-used address proxies are implemented in several different neighborhood types, including rural, suburban, small town, and urban areas. It is recognized that accessibility thresholds will vary by setting, as well as health outcome or health-related behavior. Therefore, by demonstrating how the magnitude of the distance errors can affect measures of accessibility (or exposure) to a variety of health-related spaces in different environments and at different distance thresholds, this study also makes a methodological contribution to the environmental and public health literature.

The dwelling as represented by the centroid of the building in which the study participant resides is considered the gold standard for residential address location. If dwelling centroids are not available to the researcher, then the second most accurate address proxy is the centroid of the parcel of land(i.e., the lot) on which the dwelling unit is located; this finding is true regardless of neighborhood type. When the lot centroid is used as an address proxy, accessibility misclassification errors are virtually nonexistent in urban and suburban neighborhoods, and are very minor in rural areas and small towns.

Where digital files for all residential buildings or residential lots are not available for a study region, but the researcher has access to the complete civic address (i.e., street name and number) for each subject, it is very common for researchers to geocode their tables of subject addresses using 'address locator' tools to interpolate residential addresses. While the median distance error for this address proxy is too high for researchers to simply ignore (ranging from a low of 34 m to a high of 100 m depending on facility and neighborhood types), for the most part, there are few instances of miscoded accessibility when this commonly-used address proxy is used: fewer than one-tenth (8.9%) of all observations are misclassified, except for recreation spaces within 500 m in suburban and urban neighborhoods, where approximately one-fifth of observations are misclassified (18.3% and 21.1%, respectively).

A variation on the interpolated address technique is to use the centroid of the closest street segment as address proxy. This method is useful for environmental equity studies, where researchers may want to map and visualize how access to certain environmental features varies at a fine scale across a study area, but they do not have (or cannot show for privacy reasons) specific address data for subject populations. The street segment centerline address proxy appeared to have fewer distance and misclassification errors than the more commonly-used postal code centroids, particularly for small town and rural areas.

Postal codes are certainly the most commonly-used proxy for residential addresses of research subjects in Canadian public health studies. In Canada, the postal code centroid is often the best solution when exact addresses are unavailable, or inaccessible due to research ethics board policies and privacy concerns. Our results indicate that postal code centroids are reasonably accurate proxies for residential addresses in urban and suburban areas (median positional discrepancies between 54 and 109 m depending on facility type); however, we recommend that postal codes should be used only with extreme caution for studies based in small town and rural areas of Canada. Positional discrepancies between postal code centroid and dwelling centroid can be very high in rural areas: depending on facility type, median distance errors in rural areas ranged between 762 and 1363 m. Furthermore, we found that postal codes are reasonably accurate for accessibility studies when distance thresholds are 1000 m or greater; however, we advise that postal codes should not be used as proxies for residential addresses in accessibility studies where the threshold distances or density buffers are as short as 500 m. Postal code centroids are particularly prone to misrepresenting accessibility in small towns and rural Canada, and therefore should only be used with more caution in spatial epidemiologic research in Canada.

Urban areas show the smallest distance error for all address proxies followed by suburban, small town, and rural neighborhoods. As expected, the magnitude of distance errors and threshold misclassification errors are larger, or most problematic, when the address proxy is the centroid of a large geographic aggregation such as the census tract. In general, the census tract performed poorly as an address proxy except in urban areas where threshold distances are 1600 m or greater. Similarly, we recommend that centroids of dissemination areas and weighted dissemination areas should only be used as residential address proxies in urban areas when threshold distances are set at greater than 1000 m and in suburban areas when threshold distances are set at greater than 1600 m. As for Canadian small towns, researchers should also avoid all spatially aggregated address proxies for threshold distances less than 1.6 km as the misclassification errors are consistently large, as are the distance errors. While these recommendation are based on the empirical findings related to the specific health-related facilities examined in this study, it is recognized that the positional accuracy required for spatial epidemiology research also depends on the specific exposure and health outcome under examination (e.g. spatial

accuracy is more critical for studies of exposure to air pollution than distance to nearest hospital).

This study looked at the errors in the shortest path distances from each address proxy to the closest public recreation space, junk food outlet, grocery store, school, and hospital in a full range of neighborhood types. One way in which this study differed from previous studies of positional error is that street network distances were used in the error calculation, not the relative positional errors in terms of Euclidean distances. Since a subject must use the existing street network (or pathway network) to travel from their dwelling to access the nearest park, junk food outlet, grocery store, school, or hospital, it would be inaccurate to calculate positional errors and therefore accessibility misclassification as Euclidean or 'crow fly' distances between address proxies and dwelling centroids (except where distances are too small to require use of the network). As a necessary methodological step to create baseline distance measures for comparative purposes, this study assigned health-related accessibility scores to every residential address in the study area. These individual values are at the finest scale so that, in future, they can be aggregated in any geographic frame a researcher would see fit. By creating accessibility measures to individual dwelling centroids, researchers are no longer constrained by the (often arbitrary) boundaries of blocks, postal codes, dissemination areas, census tracts, or even counties.

There is a growing trend in public health studies, particularly within the burgeoning field of 'active living research', toward the use of 'ego-centric' units (typically defined by buffers around a study participant's residence) to characterize a participant's neighborhood in order to examine the effect that local environmental factors (e.g. the mix of land uses and coverage of sidewalks) may have on health-related behaviors such as walking (e.g. Larsen et al., 2009) and outcomes such as physical activity levels (Tucker et al., 2008). The findings of this study have revealed that if commonly-used proxies such as centroids of census tracts, dissemination areas, and even postal codes, are used instead of exact addresses, distance errors can be significantly large. If distance errors are large, such 'ego-centric' neighborhood units will be significantly 'off center' and local environments can be mischaracterized. For example, the chances of misclassifying a health-promoting feature of the neighborhood such as a park (or a health-damaging feature such as a junk food outlet) as accessible (or not) can be unacceptably high, particularly when threshold distances are short, such as the commonly-used 500 m buffer (or 5-min walk zone). If positional discrepancies are too large, it will be impossible for the researcher to resolve whether any health effects of an environment are truly present. Improving the accuracy of our distance calculations increases the utility of our findings for making decisions and enacting policies aimed at improving a population's spatial accessibility to environmental features that contribute to their overall health and well-being.

Appendix A

Table A1
Distance errors (m) from address proxy to closest junk food retailer.

Neighborhood type	% (N = 104,024)	Lot (n_p = 104,024)	Geocoded point (n_p = 104,024)	Street segment (n_p = 19,365)	Postal code (n_p = 14,265)	DB (n_p = 4210)	Weighted DA (n_p = 665)	DA (n_p = 665)	Census tract* (n_p = 94)
Rural	Mean	69	163	274	1344	984	1325	1415	1427
(n = 16,686)	Median	49	85	175	762	678	897	1054	930
	75th	74	168	370	2040	1431	1930	2033	2159
	90th	166	337	597	3742	2312	3219	3261	3473
	95th	182	471	772	4436	2835	4097	4060	4136
	99th	364	1683	1683	5832	4053	5536	5690	5383
Small town	Mean	38	69	89	1241	455	562	979	1883
(n = 14,139)	Median	29	48	65	373	146	279	509	1414
	75th	35	78	111	1786	458	623	1227	3280
	90th	56	148	181	4467	1231	1207	2528	4190
	95th	99	196	245	5099	2515	2774	3765	4791
	99th	187	351	475	6483	3418	3729	5926	5448
Suburban	Mean	12	83	111	107	186	226	250	297
(n = 54,579)	Median	9	51	75	78	126	168	176	243
	75th	11	76	125	133	255	312	334	423
	90th	17	147	238	224	430	501	564	626
	95th	35	331	380	331	558	637	730	767
	99th	167	551	625	547	881	975	1216	1037
Urban	Mean	13	51	66	71	108	126	139	195
(n = 18620)	Median	8	38	52	54	77	100	113	160
	75th	12	51	81	90	146	176	194	281
	90th	17	77	120	139	230	260	284	405
	95th	30	137	166	187	309	322	351	492
	99th	61	366	377	413	530	527	550	651

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of dwelling centroids; n – number of dwelling centroids by neighborhood type; n_p – number of address proxies.

* Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

Table A2
Distance errors (m) from address proxy to closest public recreation place.

Neighborhood type	% (N = 104,024)	Lot (n_p = 104,024)	Geocoded point (n_p = 104,024)	Street segment (n_p = 19,365)	Postal code (n_p = 14,265)	DB (n_p = 4210)	Weighted DA (n_p = 665)	DA (n_p = 665)	Census tract* (n_p = 94)
Rural	Mean	63	156	270	1645	972	1491	1520	1961
(n = 16,686)	Median	43	77	185	896	677	988	1070	1347
	75th	72	161	401	2393	1427	2177	2180	2749
	90th	158	386	608	4206	2324	3612	3561	4629
	95th	185	606	781	5458	2879	4570	4401	6017
	99th	346	1069	1118	8931	4024	6495	6097	8579
Small town	Mean	41	56	77	1779	503	645	1105	2020
(n = 14,139)	Median	38	34	52	1177	156	296	599	1723
	75th	43	60	92	3109	482	712	1513	3172
	90th	55	109	155	4076	1590	1699	2882	3768
	95th	99	175	235	5095	2770	2971	4010	6521
	99th	195	464	517	9996	3327	4521	6495	7828
Suburban	Mean	11	191	214	211	266	319	347	525
(n = 54,579)	Median	8	75	106	114	176	228	241	352
	75th	9	238	265	257	367	443	473	645
	90th	16	557	586	558	632	732	772	1031
	95th	33	745	766	761	822	920	985	1420
	99th	161	1207	1243	1231	1242	1383	1674	4993
Urban	Mean	11	182	193	195	208	242	265	293
(n = 18,620)	Median	8	84	102	109	145	185	207	247
	75th	12	257	275	279	290	347	377	419
	90th	18	513	527	518	483	523	567	593
	95th	24	632	639	639	608	638	690	714
	99th	60	937	921	953	938	1055	1084	943

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of dwelling centroids; n – number of dwelling centroids by neighborhood type; n_p – number of address proxies.
* Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

Table A3
Distance errors (m) from address proxy to closest grocery store.

Neighborhood type	% (N = 104024)	Lot (n_p = 104,024)	Geocoded point (n_p = 104,024)	Street segment (n_p = 19,365)	Postal code (n_p = 14,265)	DB (n_p = 4210)	Weighted DA (n_p = 665)	DA (n_p = 665)	Census tract* (n_p = 94)
Rural	Mean	64	281	377	2000	1095	1707	1721	2581
(n = 16,686)	Median	43	100	197	1196	810	1193	1263	1704
	75th	74	212	450	2793	1599	2476	2463	3707
	90th	168	568	805	4798	2531	4029	3877	6123
	95th	191	1420	1361	6736	2976	5102	4773	7361
	99th	380	2740	2762	11412	4122	7154	6604	9584
Small town	Mean	3	115	135	2000	471	651	1102	2730
(n = 14139)	Median	25	82	95	494	169	335	559	1870
	75th	31	121	152	3532	493	765	1501	3653
	90th	53	211	288	5529	1465	1623	2963	6821
	95th	95	454	482	8523	2234	2367	3683	9253
	99th	184	567	647	10709	3027	4722	6662	10225
Suburban	Mean	12	168	197	190	271	327	345	573
(n = 54579)	Median	6	80	100	98	141	198	201	373
	75th	9	116	157	162	294	394	404	697
	90th	16	171	258	257	614	727	762	1136
	95th	34	609	736	629	994	1147	1354	1817
	99th	164	2212	2405	2237	2190	2094	2358	3819
Urban	Mean	11	115	129	132	177	203	217	381
(n = 18620)	Median	9	59	76	79	112	145	158	343
	75th	14	88	118	129	209	262	281	553
	90th	19	232	247	274	423	442	476	752
	95th	23	587	594	580	671	656	686	871
	99th	61	854	892	902	924	935	951	1089

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of dwelling centroids; n – number of dwelling centroids by neighborhood type; n_p – number of address proxies.
* Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

Table A4
Distance errors (m) from address proxy to closest school.

Neighborhood type	% (N = 104,024)	Lot (n_p = 104,024)	Geocoded point (n_p = 104,024)	Street segment (n_p = 19,365)	Postal code (n_p = 14265)	DB (n_p = 4210)	Weighted DA (n_p = 665)	DA (n_p = 665)	Census tract* (n_p = 94)
Rural	Mean	68	147	254	1547	974	1564	1595	1850
(n = 16,686)	Median	50	94	173	913	665	1017	1140	1268
	75th	76	159	367	2339	1388	2300	2299	2616
	90th	163	294	590	3957	2284	3852	3784	4441
	95th	187	413	743	5021	2929	4795	4752	5550
	99th	378	1074	1071	7693	4060	6308	6303	7493
Small town	Mean	34	65	87	1522	445	666	1087	2155
(n = 14,139)	Median	32	51	66	711	148	361	573	1679
	75th	38	79	108	2465	477	806	1423	2954
	90th	61	115	170	4048	1311	1517	2604	5926
	95th	100	163	228	5922	2271	2723	4322	6875
	99th	189	358	483	6990	3047	4187	6560	7758
Suburban	Mean	13	82	108	109	215	272	300	510
(n = 54,579)	Median	6	60	80	82	133	187	194	363
	75th	10	84	125	136	273	357	379	667
	90th	15	126	191	206	513	609	671	1057
	95th	34	180	286	277	716	838	976	1567
	99th	166	687	713	698	1200	1341	1597	2830
Urban	Mean	13	68	81	84	139	162	171	296
(n = 18,620)	Median	9	55	66	68	101	132	140	251
	75th	14	74	100	110	186	227	241	442
	90th	19	111	147	164	295	331	349	624
	95th	23	170	190	210	387	405	426	724
	99th	61	381	417	409	654	641	651	869

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of dwelling centroids; n – number of dwelling centroids by neighborhood type; n_p – number of address proxies.
* Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

Table A5
Distance errors (m) from address proxy to closest hospital.

Neighborhood type	% (N = 104,024)	Lot (n_p = 104,024)	Geocoded point (n_p = 104,024)	Street segment (n_p = 19,365)	Postal code (n_p = 14,265)	DB (n_p = 4210)	Weighted DA (n_p = 665)	DA (n_p = 665)	Census tract* (n_p = 94)
Rural	Mean	66	176	278	2382	1082	1903	1854	3285
(n = 16,686)	Median	46	85	187	1363	769	1350	1255	2088
	75th	72	284	426	3683	1561	2732	2700	5223
	90th	156	458	655	6150	2508	4496	4400	7815
	95th	180	553	817	8116	3052	5708	5535	9735
	99th	359	859	1148	11812	4375	8419	8292	13483
Small town	Mean	34	178	192	1296	546	674	998	2413
(n = 14,139)	Median	27	65	100	537	349	415	538	2166
	75th	33	335	341	1589	645	832	1273	3266
	90th	56	443	450	3664	1355	1580	2373	5281
	95th	96	511	516	4320	2203	2319	3766	6095
	99th	185	821	828	8095	3060	3690	6689	9435
Suburban	Mean	12	68	93	102	255	287	301	651
(n = 54,579)	Median	5	37	67	78	176	203	204	445
	75th	9	75	127	143	326	384	390	797
	90th	16	178	189	214	556	640	647	1256
	95th	33	188	231	267	777	848	885	1689
	99th	164	367	503	441	1358	1389	1620	5312
Urban	Mean	11	101	104	114	190	207	214	414
(n = 18,620)	Median	8	75	93	101	160	166	171	343
	75th	12	181	170	175	262	292	301	580
	90th	17	193	204	225	380	434	445	835
	95th	22	200	226	263	464	538	555	1078
	99th	58	312	319	362	738	774	814	1668

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of dwelling centroids; n – number of dwelling centroids by neighborhood type; n_p – number of address proxies.
* Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

References

Acharya A, Nyirenda J, Higgs G, Bloomfield M, Cruz-Flores S, Connor L, et al. Distance from home to hospital and thrombolytic utilization for acute ischemic stroke. J Stroke Cerebrovasc Dis 2011;20(4):295–301.

Anselin L. How (not) to lie with spatial statistics. Am J Prev Med 2006;30(2):S3–6.

Apparicio P, Cloutier M, Shearmur R. The case of Montréal's missing food deserts: evaluation of accessibility to food supermarkets. Int J Health Geog 2007;6(4):12.

Apparicio P, Abdelmajid M, Riva M, Shearmur R. Comparing alternative approaches to measuring the geographical accessibility of urban health services: distance types and aggregation-error issues. Int J Health Geogr 2008;7(7).

Austin S, Melly S, Sanchez B, Patel A, Buka S, Gortmaker A. Clustering of fast-food restaurants around schools: a novel application of spatial statistics to the study of food environments. Am J Public Health 2005;95(9):1575–81.

Bjork J, Albin M, Grahn P, Jacobsson H, Ardo J, Wadbro J, et al. Recreational values of the natural environment in relation to neighbourhood satisfaction, physical activity, obesity and wellbeing. J Epidemiol Community Health 2008;62(2).

Bonner M, Daikwon H, Nie J, Rogerson P, Vena J, Freudenheim J. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology 2003;14:408–12.

Bow C, Jennifer D, Waters N, Faris P, Seidel J, Galbraith D, et al. Accuracy of city postal code coordinates as a proxy for location of residence. Int J Health Geogr 2004;3(5).

Brownson R, Hoehner C, Day K, Forsyth A, Sallis J. Measuring the built environment for physical activity: state of the science. Am J Prev Med 2009;36(S4):S99–S123.

Cayo M, Talbot T. Positional error in automated geocoding of residential addresses. Int J Health Geogr 2003;2(10).

City of London. Parcels, buildings, address points, and health facilities GIS files [DVD]. London (ON): Geomatics Division; 2010.

Cudnick M, Schmicke R, Vaillancourt C, Newgard C, Christenson J, Davis, et al. A geospatial assessment of transport distance and survival to discharge in out of hospital cardiac arrest patients: Implications for resuscitation centers. Resuscitation 2010;81:518–23.

DMTI Spatial Inc. Database of postal code centroids and street centerline GIS files [Internet], Ottawa(On);2009. Available from <http://equinox.uwo.ca>.

Fotheringham S. Scale-independent spatial analysis. In: Goodchild M, Gopal S, editors. Accuracy of spatial data. London: Taylor & Francis; 1989. p. 221–8.

Gilliland J. The Built environment and obesity: trimming waistlines through neighbourhood design. In: Bunting, Filion, Walker, editors. Canadian cities in transition. 4th ed. Oxford Univ Press; 2010. p. 391–410.

Goldberg D. A Geocoding Best Practices Guide. Springfield, IL North Am Assoc Cent Cancer Registries;2008.

Henry K, Boscoe F. Estimating the accuracy of geographical imputation. Int J Health Geogr 2008;7(3).

Jacquez G, Rommel R. Local indicators of geocoding accuracy (LIGA): theory and application. Int J Health Geogr 2009;8(60).

Jones A, Bentham G, Horwell C. Health service accessibility and deaths from asthma in 401 local authority districts in England and Wales, 1988–92. Thorax 1997;52:218–22.

Larsen K, Gilliland J. Mapping the evolution of 'food deserts' in a Canadian city: supermarket accessibility in London, Ontario, 1961–2005. Int J Health Geogr 2008;7(16).

Larsen K, Gilliland J, Hess P, Tucker P, Irwin J, He M. The influence of the physical environment and sociodemographic characteristics on children's mode of travel to and from school. Am J Public Health 2009;99(3):520–6.

Lee R, Cubbin C, Winkleby M. Contribution of neighbourhood socioeconomic status and physical activity resources to physical activity among women. J Epidemiol Community Health 2007;61: 882–90.

Maroko A, Maantay J, Sohler N, Grady K, Arno P. The complexities of measuring access to parks and physical activity sites in New York city: a quantitative and qualitative approach. Int J Health Geogr 2009;8(34).

Middlesex County. Database of parcels, address point, aerial photos, and health facilities GIS files [DVD]. London (ON): Middlesex County Planning Dept.;2011.

Middlesex-London Health Unit. Database of food retailers [DVD].London (ON): Middlesex County Food Inspection Dept.;2010.

Morland K, Evenson K. Obesity prevalence and the local food environment. Health Place 2009;15:491–5.

Muller S, Tscharaktschiew S, Haase K. Travel-to-school mode choice modelling and patterns of school choice in urban areas. J Transport Geog 2008;16:342–57.

Nicholl J, West J, Goodacre S, Turner J. The relationship between distance to hospital and patient mortality in emergencies: an observational study. Emerg Med J 2007;24:665–8.

Panter J, Jones A, van Sluijs E, Griffin S. Attitudes, social support and environmental perceptions as predictors of active commuting behaviour in school children. J Epidemiol Community Health 2009;61:389–95.

Pearce J, Hiscok R, Blakely T, Witten K. The contextual effects of neighbourhood access to supermarkets and convenience stores on individual fruit and vegetable consumption. J Epidemiol Community Health 2008;62:198–201.

Rushton G, Armstrong M, Gittler J, Greene B, Pavlik C, West M, Zimmerman D. Geocoding in Cancer Research. Am J Prev Med 2006;30(2):S16–24.

Sadler R, Gilliland J, Arku G. An application of the edge effect in measuring accessibility to multiple food retailer types in Southwestern Ontario, Canada. Int J Health Geogr 2011;10:34.

Sarmiento OL, Schmid TL, Parra DC, Diaz-del-Castillo A, Gomez LF, Pratt M, Jacoby E, Pinzon JD, Duperly J. Quality of life, physical activity, and built environment characteristics among Columbian adults. J Phys Act Health 2010;2010 7(S2):S181–95.

Schlossberg M, Greene J, Phillips P, Johnson B, Barker B. School trips: effects of urban form and distance on travel mode. Am Plann Assoc: J Am Plann Assoc 2006;72(3):337–46.

Schootman M, Sterling D, Struthersa J, Yan Y, Laboubea T, Emo B, et al. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. Ann Epidemiol 2007;17(6):464–70.

Sharkey J. Measuring potential access to food stores and food-service places in rural areas in the US. Am J Prev Med 2009;36(4):S151–5.

Statistics Canada. Census boundary files [Internet]. Ottawa (On); Data Liberation Initiative;c2006. Available from <http://equinox.uwo.ca>.

Statistics Canada. Rural and Small Town Canada Analysis Bulletin 2011. Available from <http://www.statcan.gc.ca/pub/21-006-x/21-006-x2001003-eng.pdf>.

Strickland M, Siffel C, Gardner B, Berzen A, Correa A. Quantifying geocode location error using GIS methods. Environ Health 2007;6:10.

Talen E. Neighborhoods as service providers: a methodology for evaluating pedestrian access. Environ Plann B Plann Des 2003;30(2):181–200.

Tucker P, Irwin J, Gilllilland J, Larsen K, He M, Hess P. Environmental influences on physical activity levels in youth. Health Place 2008;15(1):357–63.

Wang M, Kim S, Gonzalez A, MacLeod K, Winkleby M. Socioeconomic and food-related physical characteristics of the neighbourhood environment are associated with body mass index. J Epidemiol Community Health 2007;61:491–8.

Ward M, Nuckols J, Giglierano J, Bonner M, Wolter C, Airola M, et al. Positional accuracy of two methods of geocoding. Epidemiology 2005;16(4):542–7.

Wolch J, Jerrett M, Reynolds K, McConnell R, Chang R, Dahmann N, Brady K, Gilliland F, Su JG, Berhane K. Childhood obesity and proximity to parks and recreational resources: a longitudinal cohort study. Health Place 2010;17(1):207–14.

Zandbergen P, Green J. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. Environ Health Perspect 2007;115(9):1363–70.

Zandbergen P. A comparison of address point, parcel and street geocoding techniques. Comput Environ Urban Syst 2008;32(3):214–32.