Research

Establishing the soundness of administrative spatial units for operationalising the active living potential of residential environments: an exemplar for designing optimal zones

Mylène Riva*[1,2,3,4,5], Philippe Apparicio[4,3], Lise Gauvin[1,2,3] and Jean-Marc Brodeur[1,2,6]

Address: [1]Département de médecine sociale et préventive, Université de Montréal, Montréal, Canada, [2]GRIS, Groupe de recherche interdisciplinaire en santé, Université de Montréal, Montréal, Canada, [3]Centre de recherche Léa-Roback sur les inégalités sociales de santé de Montréal, Université de Montréal, Montréal, Canada, [4]Institut National de la Recherche Scientifique – Urbanisation, Culture et Société, Montréal, Canada, [5]Institute of Hazard and Risk Research, Department of Geography, Durham University, UK and [6]Analyse et évaluation des interventions en santé, AnÉIS, Université de Montréal, Montréal, Canada

Email: Mylène Riva* - mylene.riva@umontreal.ca; Philippe Apparicio - philippe.apparicio@ucs.inrs.ca; Lise Gauvin - lise.gauvin.2@umontreal.ca; Jean-Marc Brodeur - jean-marc.brodeur@umontreal.ca

* Corresponding author

Abstract

Background: In health and place research, definitions of areas, area characteristics, and health outcomes should ideally be coherent with one another. Yet current approaches for delimiting areas mostly rely on spatial units "of convenience" such as census tracts. These areas may be homogeneous along socioeconomic conditions but heterogeneous along other environmental characteristics. This heterogeneity can lead to biased measurement of environment characteristics and misestimation of area effects on health. The objective of this study was to assess the soundness of census tracts as units of analysis for measuring the active living potential of environments, hypothesised to be associated with walking.

Results: Starting with data at the smallest census area level available, zones homogeneous along three indicators of active living potential, i.e. population density, land use mix, and accessibility to services were designed. Delimitation of zones ensued from statistical clustering of the smallest areas into seven clusters or "types of environment". Mapping of clusters into a GIS led to the delineation of 898 zones characterised by one of seven types of environment, corresponding to different levels of active living potential. Homogeneity of census tracts along indicators of active living potential varied. A greater proportion (83%) of variation in accessibility to services was attributable to differences between census tracts suggesting within-tract homogeneity along this variable. However, census tracts were heterogeneous with respect to population density and land use mix where a greater proportion of the variation was attributable to within-tract differences. About 55% of tracts were characterised by a combination of three or more "types of environment" suggesting substantial within-tract heterogeneity in the active living potential of environments.

Conclusion: Soundness of census tracts for measuring active living potential may be limited. Measuring active living potential with error may lead to misestimation of associations with walking, therefore limiting the correctness of inference about area effects on walking. Future studies should aim to determine homogeneity of spatial units "of convenience" along environment characteristics of interest prior to examining their association with health. Further evidence is needed to assess the extent of this methodological issue with other indicators of environment context relevant to other health indicators.

Background

Residential areas are proximal to everyday life and are therefore likely to influence health of local populations through the possibility they provide for leading healthy lives [1,2]. An accumulating body of research shows evidence for variation in health across residential areas and the significance of area context for explaining this variation, independently of the characteristics of individuals [3-5].

Different scales, or spatial units, may be relevant to specific contextual conditions and to specific heath outcomes [6,7], as illustrated by studies reporting varying strength and magnitude of area effects on health according to the operational definition of areas [8-15] or to contextual conditions [16-19]. Nonetheless current approaches for delimiting areas mostly rely on spatial units "of convenience" such as census tracts, boroughs, or wards [3,5]. These spatial units are certainly useful because they can easily be linked to data from censuses and other surveys that can be used for measuring contextual conditions. Also, they are often designed to be homogeneous along socioeconomic conditions of populations, thus being appropriate spatial units to operationalise the socioeconomic context of areas [20] (this may not hold for other administrative units, e.g. postal code areas which are design for postal delivery purposes and may be very heterogeneous in terms of population composition). However it is to be considered that through time, the composition of the units may change leading to modification of the socioeconomic conditions which may become more heterogeneous.

Yet, other contextual dimensions relevant for health may not be optimally defined within administrative spatial units. For example, conduciveness of areas to physical activity or geographic accessibility to health services may operate on different scales than socioeconomic factors. Operationalising relevant spatial units for studying area effects on health remains a conceptual and methodological challenge [4,5,7,21-26] giving rise to issues of validity and soundness of areal units as units of analysis [27].

Operationalising small areas: issues of validity and soundness units of analysis

Construct validity refers to whether or not the measurement instrument operationalises the concept of interest. In area effects on health research, construct validity is a matter of establishing 1) the soundness of units of analysis, i.e., whether or not area boundaries are aetiologically meaningful for studying the association between area characteristics and a given health indicator, and 2) whether or not data constitute appropriate operationalisations of exposure variables, i.e. the characteristics of areas [27]. Ideally, definitions of areas, the characteristics of

these areas, and the health outcome(s) being studied should be coherent with one another [7].

Measures of area characteristics derived from population censuses and other surveys, e.g. socioeconomic position, although easily accessible, provide only partial information on the context of areas and may in fact be endogenous to the composition of the areas as they are determined by individual characteristics of residents [28]. Collecting and measuring "true" or "integral" area data, i.e. data only measurable at the area level through procedures such as ecometrics and spatial analyses has been underscored as critical for measuring unbiased area-level variables [2,7,28,29]. Likewise, defining aetiologically meaningful areas in coherence with the specific purposes of the study, either in terms of health outcomes, characteristics of environment, or associations between the two [23,28,30,31] is important for understanding the significance of residential areas for health. Measurement errors can result if the spatial patterning of environmental characteristics does not correspond to the spatial units chosen for operationalising areas and their context [27].

Defining relevant geographic areas becomes salient in light of the modifiable areal unit problem, i.e. the fact that analytical results are sensitive to the definition of spatial units at which data are aggregated [32,33]. In other words, area effects may be observed only at certain scales, i.e. scales at which data are collected and aggregated and may vary or be absent when observed at other scales. Imposing arbitrary spatial units on a continuous spatial process, e.g. characteristics of environments, may lead to the delineation of artificial spatial patterns. In such cases, environment characteristics may be measured with error. As a result the internal validity of the study, i.e. whether or not observed associations are unbiased, may be threatened.

In addition, as per spatial autocorrelation, areas will share similar contextual conditions as a function of their proximity in space [34]. By using spatial units of convenience, it is assumed that contextual conditions within one area are different and influence health independently of conditions in neighbouring areas [4,5,21,22,24-26,35], when in fact these conditions are clustered in space. Furthermore, for any area effects to be detected there must be variation in the exposure being studied [36]. Yet the variation of environment characteristics may be smoothed out by the definition of area units used to measure them. For example, if spatial units encompass environments that are both conducive to walking and others that are less so, averaging values of conduciveness over census tracts could potentially lead to mismeasurment of exposures. Within area homogeneity along the contextual conditions under examination is thus required for minimising measurement error. Correspondingly, for inferring about area

effects on health, between area differences must be maximised: if data are collected in contiguous and heterogeneous areas, variations in both characteristics of environments and health outcomes, and their association, may be misestimated. As area effects on health have been observed to be stronger in more homogeneous areas [37,38], homogeneity of areas may thus influence the estimation of area effects and therefore the validity of conclusion.

In Figure 1, we propose a template that could be useful for establishing the soundness of spatial units "of convenience" to operationally define areas for specific research questions. For example, the template could be used to guide the decision as to whether or not census tracts are the most appropriate spatial units of analysis for measuring associations between area-level socioeconomic position (SEP) and obesity. That is, if they allow for measuring indicators of SEP without bias (and ultimately for estimating non-biased association with health outcomes) by showing homogeneity in the distribution of indicators and optimising their spatial patterning. In the methods section, we propose an approach for achieving this end. Intuitively, it can be expected that census tracts are appropriate units for undertaking such a study as they are, as mentioned above, initially designed to be homogeneous along socioeconomic conditions. But across time, the socioeconomic composition of census tracts may change as people migrate in and out of areas, potentially introducing heterogeneity in the socioeconomic make-up of

the area. This could result in a "dilution" of the true level of deprivation. Averaging indicators of SEP over census tracts thus may mask "pockets" of poverty. The exercise of establishing the soundness of census tracts as units of analysis would be important here, as it would allow to measure with less error indicators of SEP and their association with health outcomes. In multilevel studies, mismeasurement of environment characteristics may influence the strength of the observed association between environment characteristics and health indicators [39]. As such, associations may not be detected or may be spurious, therefore limiting the precision of research findings for informing public health and public policy actions to tackle social and geographical inequalities in health.

Establishing the soundness of spatial units of analysis chosen for operationalising area boundaries and measuring area context is an important methodological consideration, but it is often overlooked. Alternatively, designing spatial units of analysis maximising homogeneity of selected environment characteristics may prove to be a viable strategy for advancing the understanding of processes linking place to health [7].

Objectives
The aim of this investigation is to assess the soundness of census tracts as units of analysis for studying associations between a specific exposure and a specific health outcome, namely the active living potential of residential environments and walking behaviours. Active living potential refers to the conditions of areas that encourage the likelihood of integrating physical activity into daily routines [29]. Census tracts were selected as spatial units "of convenience" because of extensive use of this spatial unit of analysis in current research on health and place [4,5]. In Canada, census tracts are small and relatively stable geographic areas with populations ranging in size between 2500 and 8000 inhabitants; at the time of their creation, census tracts were homogeneous in terms of socioeconomic characteristics, e.g. economic status and social living conditions [40].

To establish the soundness of census tracts as a unit of analysis, we developed and tested a comprehensible method for designing optimal and homogeneous spatial units espousing the spatial distribution of selected environment characteristics linked to the concept of density of destinations that is the physical and social characteristics of residential areas related to land use pattern [29]. Three indicators were used to operationalise the construct of active living potential: population density, land use mix, and geographic accessibility to proximity services. The specific objectives of the study are to examine whether or
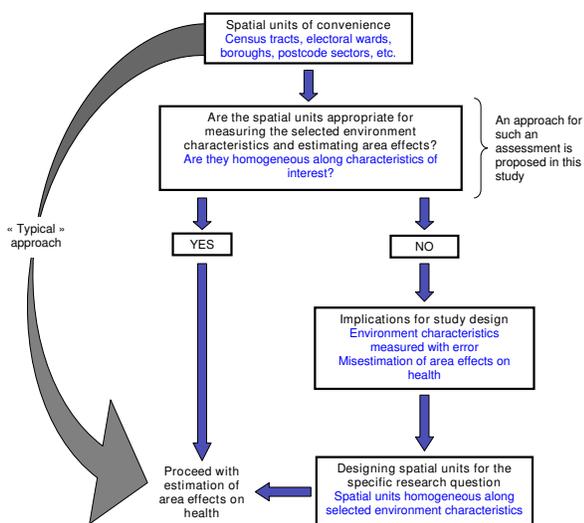

Figure 1
Template for deciding upon the soundness of "spatial units of convenience" to operationally define small area units of analysis.

not: 1) census tracts are homogeneous units of analysis along indicators of active living potential; 2) active living potential and socioeconomic indicators follow a similar spatial distribution; and 3) census tracts encompass smaller areas with different (or similar) levels of active living potential.

Active living potential was chosen because of increasing research reporting associations between this environmental construct and walking [19,41-53], an important public health indicator [54-56]. This choice was also motivated by availability of spatial datasets allowing for the operationalisation of integral measures of land use mix and geographic accessibility to services in geographical information systems, and by the availability of individual-level data on walking behaviours (to be examined in future analyses).

Methods

The methodology section includes two parts. First, we present criteria and methods for designing homogeneous areas (henceforth designated as "zones"). Second, we present analyses undertaken to assess the soundness of census tracts as units of analysis for measuring the active living potential of residential areas.

Designing optimal, homogeneous zones

Zone design refers to the placement of areal unit boundaries [9]. It can be achieved discursively (manually) by grouping basic spatial units into larger ones [57-59], by combining social, statistical, and spatial analysis methods [60,61], and automatically through computationally intensive automated zoning software [9,15,37,62-65].

Three criteria guided the choice of the method for zone design. First, we wanted to design zones based on the spatial distribution of environmental characteristics related to active living potential, namely population density, land use mix, and geographic accessibility to selected proximity services. We had no requirement regarding population and area sizes as zones were defined on the basis of the spatial distribution of these characteristics. Second, the method for zone design had to be optimal, i.e. to maximize variation between zones and to minimize variation within zones in the selected characteristics. In other words, the aim was to design zones that were internally homogeneous on the three indicators of active living potential, but different (heterogeneous) amongst themselves. Finally, we wanted a method that was rigorous but comprehensible and easy to implement. We opted for an approach that combined a statistical classification method, *K*-means clustering, to mapping applications in geographic information system. This three-step approach is described in greater details in the following sections.

Step 1: Measuring environment characteristics at the smallest area level

The study area is the Island of Montreal, Canada, an urban centre with 1 812 723 residents. As of January 2006, on the Island of Montreal, there are 15 municipalities, in addition to the municipality of Montreal which includes 19 boroughs [66]. The Island of Montreal is further divided into 521 census tracts and 3222 dissemination areas. Dissemination areas (DAs) were used as basic spatial units for designing zones because they are the smallest standard geographic areas for which Canadian census data are available (population size between 400 and 700 residents) [40]. On the Island of Montreal, their average size is 0.15 km² (ranging between 364 m² and 18 km²) with an average population of 562 individuals (ranging between 44 and 2138 residents). DA values for population density, land use mix, and accessibility to services were computed in a geographical information system (ArcGIS 9.2) [67].

Population density refers to number of individuals per unit area. It was computed by dividing the total number of residents of a DA by its area size (km²) [68].

Land use mix relates to the diversity or variety of land uses within an area. It was computed using an entropy index [47,69,70] which measures the homogeneity or diversity of land uses within a spatial unit. The index is defined as follow:

$$E_j = -\sum_{i=1}^{n}\left[\left(A_{ij}/D_j\right)\ln\left(A_{ij}/D_j\right)\right]/\ln n \qquad (1)$$

Where A_{ij} is the surface area of land use i in dissemination area j, D_j is the surface area of dissemination area j, and n is the total number of possible land uses which in the current case corresponds to 16, the number of different land uses characterising the Island of Montreal [71]. The index values range between 0 and 1, where 1 corresponds to a highly mixed area, and 0 to a homogeneous area, that is an area characterised by only one type of land use (e.g. low density housing). This index has been used in many studies to measure land use mix [47,72].

Geographic accessibility to proximity services refers to geographic distance to or from destinations, here to supermarkets, pharmacies, banks, and libraries. These services were selected because they are most likely to be used on a regular basis, conveying the idea of proximity services potentially accessible through walking. There are many measures of geographic accessibility [73,74]. In this study, geographic accessibility was defined in terms of the number of the selected services within an area, conferring the notion of the offer of services provided by the immediate surroundings. Supermarkets, pharmacies, banks, and libraries

were geocoded at the parcel level [75]. In order to minimise aggregation errors [73,76], accessibility was measured by computing distances of services located within a one kilometre (network distance) radius [77] from the centroid of census blocks (n = 14 527) comprised within any one DA; the distances were than averaged and weighted by the total population of each census blocks.

Characterisation of DAs along the three indicators resulted in a sample of 3206 DAs. Measures of land use mix and accessibility to services were normally distributed; population density was normalized using a LOG10 transformation [78]. Population density was significantly and positively correlated to accessibility to services (r = 0.45, p < 0.001), and negatively to land use mix (r = -0.32, p < 0.001). Land use mix and accessibility to services were not significantly correlated (r = 0.03, p > 0.500). Prior to cluster analyses, these variables were standardized to a mean of 0 and a standard deviation of 1, higher values representing greater levels of population density, land use mix, and accessibility to services.

Step 2: Classifying smallest areas into clusters, e.g. "types of environments", using K-means clustering

K-means statistical clustering techniques using SAS (version 9.1) for Windows [79] was applied to classify DAs into *k* number of optimal clusters homogeneous in terms of active living potential. In social sciences, notably in geography, *K*-means is largely employed to classify areas (e.g. geodemographics [80]). The method uses an allocation/re-allocation algorithm to optimally reassign objects, here DAs, to the nearest cluster centroid [81-83]. The goal is to maximize between cluster variations and to minimize within cluster variations. The aim of this second step was to group DAs with similar values of population density, land use mix, and accessibility to services into *k* types of environments that are internally homogeneous but different among them. These types of environments correspond to different levels of active living potential. For *K*-means clustering, the number of clusters (*k*) must be determined at the onset of analyses; as we had no a priori for such number, we conducted analyses for *k* = 4 to *k* = 20.

Step 3: Mapping the clusters to create optimal and homogeneous zones

In a final step, the *k* types of environments were imported into ArcGIS 9.2 and mapped out. This lead to the delineation of *n* homogeneous zones i.e., units of analysis, characterised by one of *k* active living potential.

Statistical analyses: Assessing the soundness of census tracts as units of analysis for operationalising active living potential

The soundness of census tracts for operationalising indicators of active living potential was assessed through three series of analyses.

First, to assess the homogeneity of census tracts, variation in indicators of active living potential was estimated and decomposed between and within areas. Population density, land use mix, and accessibility of services were measured continuously at the DAs level (level 1: n = 3206). In separate two-level multilevel models, DAs were nested into zones (n = 898) and into census tracts (n = 506 with valid population and socioeconomic data). Between-area variation in indicators of active living potential was estimated using the intraclass correlation coefficient (ICC) from unconditional (null) multilevel models using HLM software Version 6.04 [84]. The ICC indicates the proportion of variation in a dependent variable that is attributable to differences between area units. Greater ICC values indicate that variation of a variable is greater between units than within, i.e. units are different among them but internally homogeneous. Using the same analytical approach, homogeneity of zones and census tracts along indicators of socioeconomic position was assessed and compared. DA-level data on the proportion of low-income households, of people with less than high school education, and of people with a university degree were obtained from the 2001 Canadian census.

Second, analysis of variance was performed to examine the proportion of variation across zones in socioeconomic variables explained by the *k* types of environment. Indicators of SEP at the DA-level were aggregated (weighted by population) at the zone-level. These analyses were performed to examine whether or not socioeconomic and active living indicators follow a similar spatial distribution as is implicitly assumed when measured within the same area unit of analysis.

Finally, descriptive statistics were employed to assess the extent to which the spatial distribution of the different types of environment coincides with the boundaries of census tracts. These analyses were conducted to examine if census tracts encompassed environments with differing levels of active living potential. The numbers of zones straddling over one or more census tracts, and the number of types of environment encompassed within census tracts were computed. To examine whether or not the spatial distribution of more mixed or more homogeneous census tracts (i.e. the number of types of environments encompassed within census tracts) was structured in space, global values of spatial autocorrelation were computed using Moran *I* with a first-order contiguity matrix [85,86]. Values for Moran *I* vary between -1 and 1, where negative values indicate negative spatial autocorrelation, i.e. neighbouring spatial units have different values, and positive values indicate positive spatial autocorrelation, i.e. neighbouring units have similar values. The covariance in Moran *I* is the covariance over space for neighbouring spatial units, and will not be computed unless two units are contiguous (first order); also, only one variable is consid-

ered [85], here the number of types of environments included in census tracts.

Results
Description of types of environment and zones
Figure 2 illustrates results of the *K*-means clustering, which show that the 3206 DAs were optimally classified into 7 clusters or "types of environments" as indicated by peaks [87] in both the Pseudo-F statistic [88] and the Cubic clustering criterion [89]. These clusters explain 72.8% of the total variation in the three indicators of active living potential. Thus, differences among the seven clusters and similarity of DAs comprised within the same cluster, i.e. within-cluster homogeneity, were both maximized. The seven types of environments correspond to seven different levels of active living potential. They encompassed more suburban to more central urban types of environments defined by different values of population density, land use mix, and accessibility to services. The types of environments are described in Figure 2 and Figure 3.

Low-density and mid-density suburban areas are characterised by lower values of population density and accessi-

bility to services. Diverse central urban areas and central urban areas with high accessibility are more densely populated and have greater access to services than any other types of environment. Although population density and accessibility to services follow to some extent an increasing gradient from more suburban to more urban areas, the pattern of land use mix is more complex: there are low values in urban areas and high values in suburban areas. Dissemination areas are designed to be similar in population size (among other characteristics); thus the area size required to reach the set population threshold (i.e. between 400 to 700 residents [40]) will be larger in less densely populated areas and smaller in more urban areas. As a consequence, larger dissemination areas are more likely to encompass different land use than are smaller dissemination areas located in urban areas.

Figure 2 also presents the statistical proximity (Euclidian distance) of the centroids of clusters (cluster mean values), i.e. types of environment, in a three dimensional graph where the axes correspond to the three indicators of active living potential. With respect to their spatial distribution, the types of environment are positively correlated in space indicating that contiguous zones were characterised by similar types of environment.
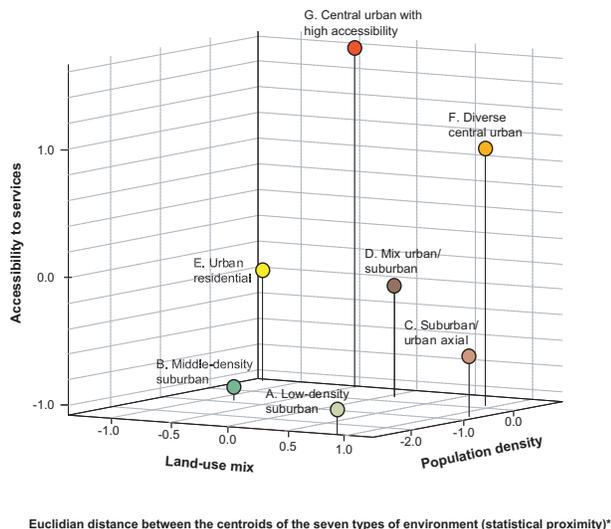
Mapping of the clusters into the GIS led to the delineation of 898 zones or units of analysis characterised by one of the seven types of environments, i.e. active living potential, as illustrated in Figure 3. Zones are significantly smaller than census tracts, an average of 0.54 km² (SD = 3.50) compared to 0.96 km² (SD = 1.98) (t = -2.46; p < 0.05), but the variation of their area size is not statistically different (F = 0.68; p = 0.409). Zones are significantly smaller than census tracts in population size, an average of 1960 (SD = 3867) residents compared to 3554 (SD = 1647) (t = -10.73; p < 0.001), and there is significantly greater variability in population size across zones than across census tracts (F = 11.40; p < 0.01). Zones characterised by more suburban contexts are on average larger and have relatively smaller population counts than urban zones.

Soundness of census tracts as units of analysis for measuring active living potential
Homogeneity of census tracts along active living potential indicators
Results of homogeneity of zones and census tracts along active living indicators appear in Figure 4. The variation in indicators is not uniform across census tracts. A greater proportion (83%) of variation in accessibility to services is attributable to differences between census tracts, as indicated by a higher ICC value, suggesting within census tract homogeneity along this indicator. Yet about half of the variation in population density is between census tracts (52%), whereas there is greater variation in land use mix

Euclidian distance between the centroids of the seven types of environment (statistical proximity)*

	A	B	C	D	E	F
A. Low-density suburban						
B. Middle-density suburban	2.53					
C. Suburban/urban axial	2.04	2.01				
D. Mix urban/suburban	3.20	1.63	1.54			
E. Urban residential	4.18	1.86	2.89	1.43		
F. Diverse central urban	3.40	2.76	1.78	1.44	2.56	
G. Central urban with high accessibility	4.51	3.04	3.12	1.94	1.94	1.68

* The lower the value, the closer the centroids.

Figure 2
Statistical proximity of the seven types of environment (clusters).

Types of environment		Zones		Surface area	
		N	%	km²	%
A	Low-density suburban	31	3.5	222.2	45.5
B	Middle-density suburban	104	11.6	75.2	15.4
C	Suburban/urban axial	146	16.3	83.7	17.1
D	Mix urban/suburban	243	27.1	41.0	8.4
E	Urban residential	212	23.6	17.1	3.5
F	Diverse central urban	85	9.5	32.6	6.7
G	Central urban with high accessibility	77	8.6	16.5	3.4
	Total	**898**	**100.0**	**488.3**	**100.0**

No data   Census tracts

Standardised mean values of indicators of active living potential for each type of environment

Population density
Land use mix
Accessibility to services

Description of types of environment

A Low-density suburban — Suburban areas characterised by lowest population density, land use diversity, and low accessibility to services.

B Middle-density suburban — Suburban areas characterised by low population density, land use homogeneity, and weak accessibility to services.

C Suburban/urban axial — Suburban and urban areas characterised by low population density, land use diversity, and weak accessibility to services. Located close to main street axis.

D Mix urban/suburban — Suburban and urban areas with average values of population density, land use mix, and accessibility to services. Located mainly in the central and eastern parts of the Island.

E Urban residential — Urban areas with highest population density, land use homogeneity, and weak accessibility to services. Located mainly in peripheral residential areas to central urban areas.

F Diverse central urban — Urban areas characterised by average population density, land use diversity, and good accessibility to services.

G Central urban with high accessibility — Urban areas with high population density, land use homogeneity, and highest accessibility to services. Located mainly in central areas of the Island.
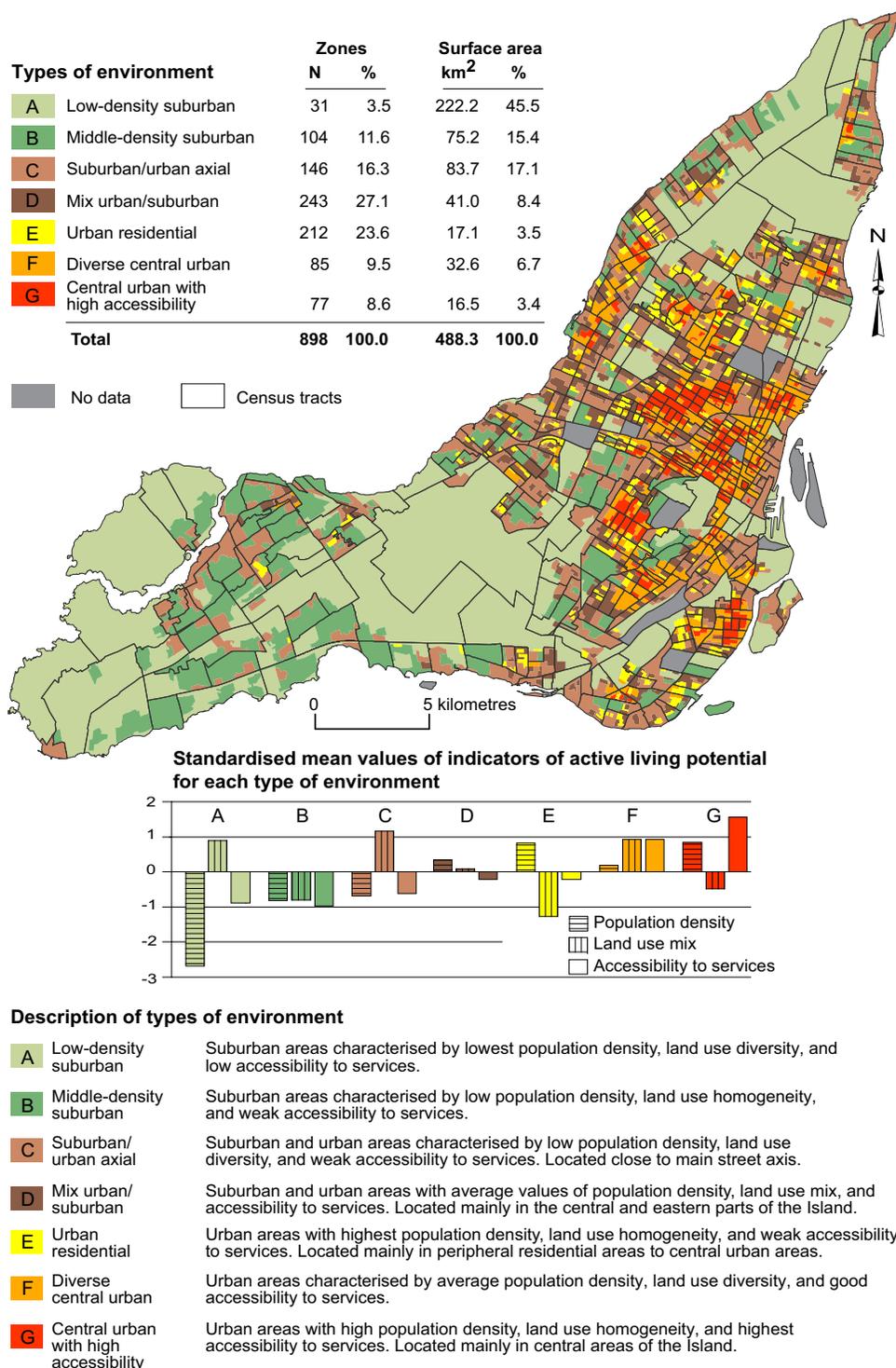
Figure 3
Description of types of environment (clusters) and zones.

within census tracts (85%), indicating greater heterogeneity of tracts along these indicators. The degree of homogeneity of tracts therefore varies according to the indicator examined. For population density and land-use mix, but not for accessibility of services, variation between zones is greater than variation between census tracts. This shows that the method was successful in designing areas or units of analysis that were more homogeneous than census tracts along dimensions of active living potential.

The degree of homogeneity of census tracts and zones along socioeconomic indicators shows that for the selected variables, variability is larger between census tracts than between zones (Figure 4). Census tracts are relatively homogeneous areas in terms of the socioeconomic environment, especially for proportion of population with a university education.

Spatial distribution of active living potential and socioeconomic indicators

Examining variation in socioeconomic indicators across zones shows that they follow a different spatial distribution than that of active living potential indicators. Results of analyses of variance (results not shown) revealed that 15.2% of the variation in the proportion of low-income households was explained by the seven types of environment whereas these proportions were 5.2% for the proportion of people with less than high school and 3.8% for the proportion of people with a university education.

Types of environments encompassed within census tracts boundaries
Overall, zones are not well contained within census tracts. As shown in Figure 5, only 30.5% of zones are completely located within the boundaries of one census tract. Forty-eight percent of zones straddle two or three census tracts whereas, 21.5% spread over more than four tracts. Corre-
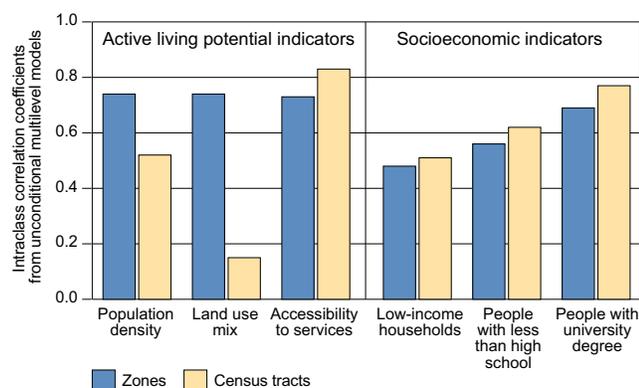
spondingly, there is considerable variability in types of environment within census tracts.

As illustrated in Figure 6, 11.2% of census tracts encompass only one type of environment and 34.3% encompass two types. About 28% of census tracts are characterised by three different types of environment, whereas 26.3% comprise 4 or more different types. Among census tracts encompassing two types of environment (n = 175), about two-thirds (66.3%) comprise types that are statistically similar as indicated by distances between their centroids (two or less distance lag as indicated in the distance matrix in Figure 2; results not shown). For example, census tracts often comprise a combination of low-density suburban and suburban/urban axial zones (26.3%), or a grouping of diverse and high accessibility central urban areas (35.4%). Globally, the number of types of environment encompass within census tracts is positively correlated in space (Moran I = 0.26; p < 0.001), suggesting that more homogeneous or more mixed census tracts are often contiguous in space (Figure 6). More heterogeneous census tracts are located mainly on the periphery of central urban areas and in the eastern part of the Island of Montreal, and to a lesser extend in the west-end suburbs.

Discussion
The objective of this study was to assess the soundness of census tracts as units of analysis, i.e. their degree of homogeneity in terms of the active living potential of residential environments associated with walking. In order to do so, homogeneous zones that optimised the spatial patterning of active living potential indicators hypothesised to be associated with greater involvement in walking, namely population density, land-mix use, and accessibility to services, were successfully designed. This was done through the application of an easy-to-use method combining a classification method called *K*-means clustering with basic mapping applications of geographical informa-
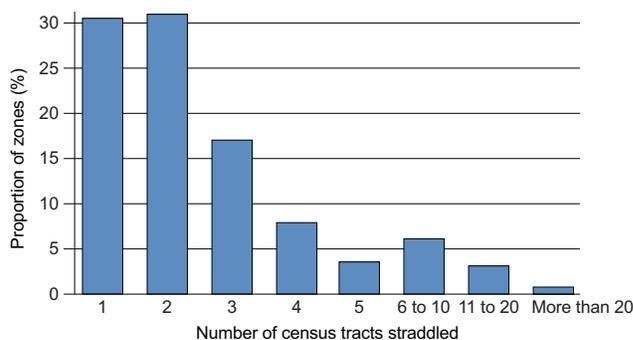
Figure 4
Decomposition of variation in indicators of active living potential and socioeconomic position across zones and across census tracts.

Figure 5
Proportion of zones straddling different numbers of census tracts across the Island of Montreal.
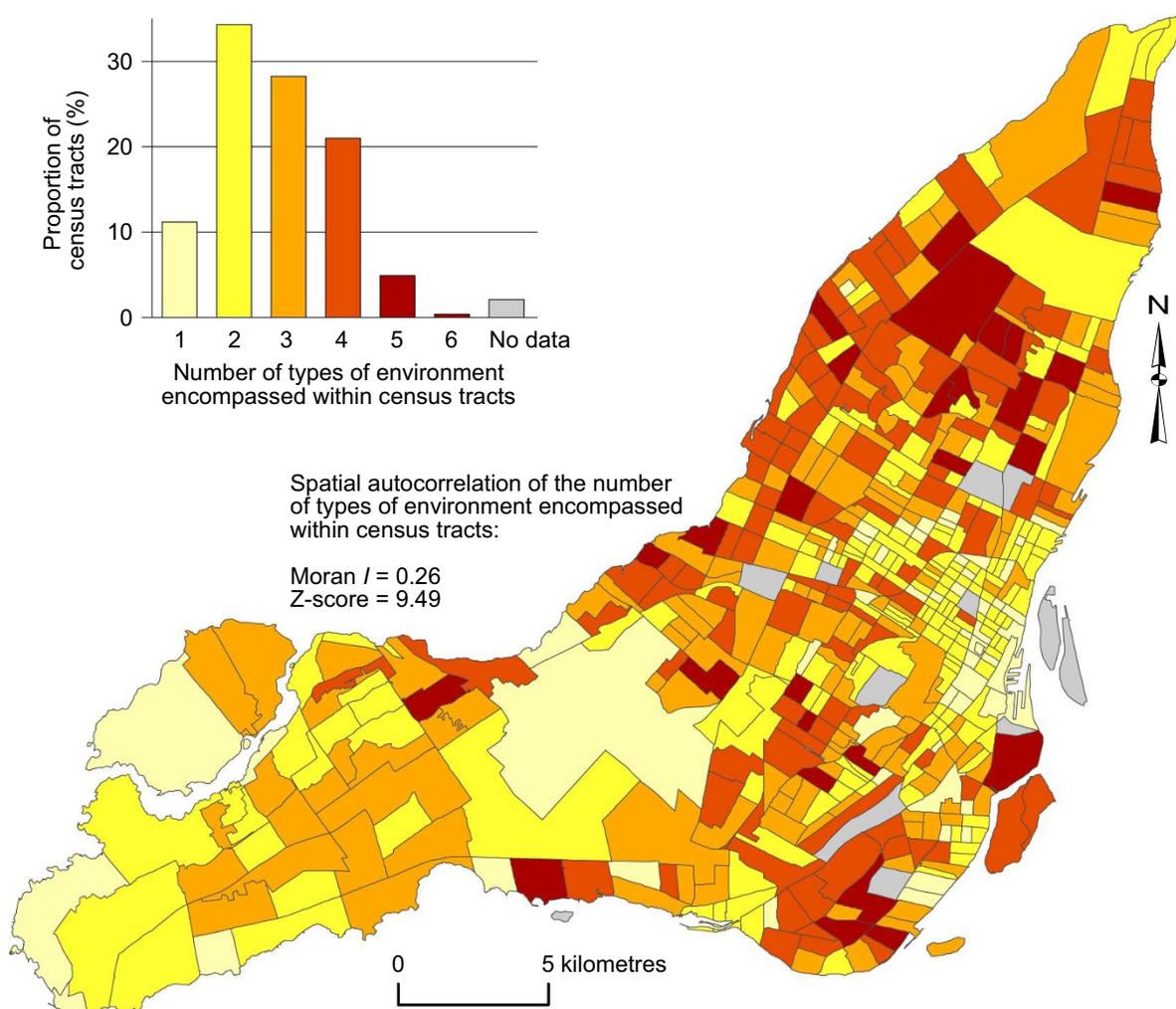
Figure 6
Number of types of environment encompassed within census tract boundaries.

tion systems. The degree of soundness of census tracts as units of analysis was established through a series of analyses comparing them to the newly-designed zones.

First the distribution of the three active living indicators between and within census tracts was assessed. Although census tracts were homogeneous in terms of accessibility to services, they were less homogenous in population density; for this indicator within and between census tracts variations were about equal. Census tracts were clearly not homogeneous in terms of land use mix as the variability within tracts largely exceeded the variability between tracts. In contrast, census tracts were homogeneous along socioeconomic variables. These results suggest

that the spatial patterning of the active living potential of environments do not neatly follow in the delineation of census tracts, which may be more suitable as units of analysis for operationalising socioeconomic contexts.

Then, findings revealed that the spatial distribution of active living and socioeconomic indicators followed different spatial distribution. At the zone-level, types of environment explained a small proportion of variation of socioeconomic variables. This indicates that processes underlying the distribution of active living and SEP indicators, although potentially linked [2,6], operate at different scales and thus require different units of analysis.

In the final set of analyses, within tract variability in terms of what we labelled "types of environment" was examined. This allowed for the assessment of whether or not census tracts encompassed environments that were substantively different among them in terms of their active living potential. Census tracts comprising two different types of environments (34.3%) were not considered necessarily as problematic, given that some types of environment were more similar than others and were often contiguous in space. For example, diverse and high accessibility central urban zones were often contiguous in space and were statistically most similar (as indicated by statistical distances between clusters; Figure 2). However, census tracts comprising three or more types of environment raised concerns; such a situation was observed in more than half of census tracts. These tracts encompass environments that are simultaneously most conducive to walking and others that are least so. Averaging values of conduciveness to walking could potentially lead to significant errors when measuring active living potential at the census tract-level.

The approach for defining areas or units of analysis differs from those involving the definition of strictly "ecologically meaningful" or "natural" neighbourhoods, i.e. neighbourhoods imbued with meaning for residents [21] or as consisting of a group of homes sharing a commonly defined residential area often having name [20]. Defining such units of analysis is important when the notion of commonly shared territory is related to the contextual condition of interest, for example social capital or collective efficacy [7,27]; this notion is not conjured up by active living potential. Designing zones based on the spatial distribution of active living indicators empirically linked to greater involvement in walking leads to the definition of areas that are more appropriate units of analysis and increases the internal validity of study design examining the environmental determinants of walking.

Future studies are needed to assess the impact of the choice of other environmental characteristics for designing zones relevant to other health indicators, and to other geographical areas. For example, areas relevant for studying the social and environmental determinants of overweight and obesity may be delimited according to the distribution of active living variables and food provision (accessibility of both healthy and non-healthy food). For studying mental health outcomes, social dimensions of area context such as social support and opportunities for social participation, may be more relevant. It is to be expected that designing zones using other indicators of contextual conditions associated with other health outcomes will lead to different spatial configuration of area units of analysis.

Homogeneous zones are designed with the aim of optimising the study of a phenomenon or for the purpose of uncovering the aetiology underlying associations between area context and health. As such, the configuration of zones should not be viewed as other "spaces" of actions for public health and policy interventions. Rather, they may be useful for informing on viable interventions and policy strategies that may be health promoting.

Limitations
Results of this study should be considered in light of some limitations. First, there is a seven year time lag (2000 to 2006) between the dates of creation of the different datasets used to characterise dissemination areas in terms of their active living potential and socioeconomic position. Although changes in the built environment may have taken place during this period, the speed at which changes occur is not well documented; however over a seven-year period, changes in the built environment can be expected to be modest.

Other indicators of active living potential could be examined in designing homogeneous areas, such as street connectivity, safety, and accessibility to other services or resources such as parks. In this study, the measurement of land use mix was dependent on the size of disseminations areas which are defined in part by a population size threshold: because of lower population density in suburban areas, DAs are likely to span a greater territory and therefore encompass more types of land use. Other scales for measuring land use mix could be considered [47].

Conclusion
For studies concerned with the social and environmental determinants of health and more specifically of physical activity, results of this study have several implications. Delimiting areas is a key conceptual and methodological challenge in research on health and place. In this paper, we developed an easy-to-use method for establishing homogeneous units of analysis in terms of specific environmental characteristics hypothesised to be linked to a specific health indicator. The focus was on active living potential of areas and walking behaviours. Using these homogeneous zones as comparison, the objective was to assess the soundness of spatial units "of convenience", i.e. census tracts, to operationalise contexts for which they were not purposely developed. The methods developed in this study add to the growing literature on alternative ways to conceptualise and define the boundaries of area units for studying the determinants of health.

Findings showed that although census tracts may be homogeneous along independent indicators of active living potential, they were most often characterised by a combination of types of environment that were substan-

tively different in terms of their active living potential. For this reason, census tracts should be used with caution as units of analysis when operationalising active living potential for studying determinants of walking. But census tracts or other administratively defined areas may be appropriate area units, i.e. may be homogeneous enough, when processes hypothesised to be operating on health are linked to the socioeconomic context of an area, for example affluence or poverty.

In this study, zones were delimited for methodological and aetiological purposes with the aim of minimising measurement errors of environmental characteristics and increasing internal validity of study design for measuring area effects on health. As can be expected, the zones are context-specific and cannot be exported to other geographic areas. Rather they are representations of the local realities of processes relating environmental characteristics to health. As suggested by others, the geographical aspects of the study design should be considered prior to conducting analyses [15]. Establishing the soundness of spatial units "of convenience" for representing the environmental and spatial processes under investigation should be part of the empirical approach for conceptualising, operationalising, and measuring area effects on health.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MR conceptualised the study. She carried out spatial and statistical analyses, mapping of results, and drafted the manuscript. PA, LG and JMB participated in the conceptualisation the study, and in data analyses. All authors critically revised the paper, and read and approved the final manuscript.

Acknowledgements

References

1. Curtis S, Jones IR: **Is there a place for geography in the analysis of health inequality?** *Sociology of Health & Illness* 1998, **20(5):**645-672.
2. Macintyre S, Ellaway A, Cummins S: **Place effects on health: how can we conceptualise, operationalise and measure them?** *Social Science & Medicine* 2002, **55(1):**125-139.
3. Ellen IG, Mijanovich T, Dillman K-N: **Neighborhood Effects on Health: Exploring the Links and Assessing the Evidence.** *Journal of Urban Affairs* 2001, **23(3&4):**391-408.
4. Pickett KE, Pearl M: **Multilevel analyses of neighbourhood socioeconomic context and health outcomes: A critical review.** *Journal of Epidemiology and Community Health* 2001, **55(2):**111-122.
5. Riva M, Gauvin L, Barnett TA: **Toward the next generation of research into small area effects on health: A synthesis of multilevel investigations published since July 1998.** *Journal of Epidemiology and Community Health* 2007, **61(10):**853-861.
6. Daniel M, Moore S, Kestens Y: **Framing the biosocial pathways underlying associations between place and cardiometabolic disease.** *Health and Place* 2008, **14(2):**117-132.
7. Diez Roux AV: **Investigating neighborhood and area effects on health.** *American Journal of Public Health* 2001, **91(11):**1783-1789.
8. Blakely TA, Lochner K, Kawachi I: **Metropolitan area inequality and self-rated health – a multi-level study.** *Social Science & Medicine* 2002, **54(1):**65-77.
9. Cockings S, Martin D: **Zone design for environment and health studies using pre-aggregated data.** *Social Science & Medicine* 2005, **60(12):**2729-2742.
10. Franzini L, Spears W: **Contributions of social context to inequalities in years of life lost to heart disease in Texas, USA.** *Social Science & Medicine* 2003, **57(10):**1847-1861.
11. Hou F, Myles J: **Neighbourhood inequality, neighbourhood affluence and population health.** *Social Science & Medicine* 2005, **60(7):**1557-1569.
12. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R: **Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: Does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project.** *American Journal of Epidemiology* 2002, **156(5):**471-482.
13. Reijneveld SA, Verheij RA, de Bakker DH: **The impact of area deprivation on differences in health: does the choice of the geographical classification matter?** *Journal of Epidemiology and Community Health* 2000, **54(4):**306-313.
14. Oliver L, Hayes MV: **Does choice of spatial unit matter for estimating small area disparities in health and place effects in the Vancouver census metropolitain area?** *Canadian Journal of Public Health* 2007, **98:**S27-S34.
15. Flowerdew R, Manley DJ, Sabel CE: **Neighbourhood effects on health: Does it matter where you draw the boundaries?** *Social Science & Medicine* 2008, **66:**1241-1255.
16. Chaix B, Chauvin P: **Tobacco and alcohol consumption, sedentary lifestyle and overweightness in France: A multilevel analysis of individual and area-level determinants.** *European Journal of Epidemiology* 2003, **18(6):**531-538.
17. Chaix B, Guilbert P, Chauvin P: **A multilevel analysis of tobacco use and tobacco consumption levels in France – Are there any combination risk groups?** *European Journal of Public Health* 2004, **14(2):**186-190.
18. Chuang YC, Cubbin C, Ahn D, Winkleby MA: **Effects of neighbourhood socioeconomic status and convenience store concentration on individual level smoking.** *Journal of Epidemiology and Community Health* 2005, **59(7):**568-573.
19. van Lenthe FJ, Brug J, Mackenbach JP: **Neighbourhood inequalities in physical inactivity: the role of neighbourhood attractiveness, proximity to local facilities and safety in the Netherlands.** *Social Science & Medicine* 2005, **60(4):**763-775.
20. Ross NA, Tremblay S, Graham K: **Neighbourhood influences on health in Montreal, Canada.** *Social Science & Medicine* 2004, **59(7):**1485-1494.
21. Coulton CJ, Korbin J, Chan T, Su M: **Mapping residents' perceptions of neighborhood boundaries: A methodological note.** *American Journal of Community Psychology* 2001, **29(2):**371-383.
22. Diez-Roux AV: **Multilevel analysis in public health research.** *Annual Review of Public Health* 2000, **21:**171-192.
23. Gauvin L, Robitaille E, Riva M, McLaren L, Dassa C, Potvin L: **Conceptualizing and operationalizing neighbourhoods: the conundrum of identifying territorial units.** *Canadian Journal of Public Health* 2007, **98(suppl 1):**S18-S26.

24. O'Campo P: **Invited commentary: Advancing theory and methods for multilevel models of residential neighborhoods and health.** *American Journal of Epidemiology* 2003, **157(1):**9-13.
25. Subramanian S, Jones K, Duncan C: **Multilevel methods for public health research.** In *Neighborhoods and health* Edited by: Kawachi I BL. New York: Oxford University Press; 2003:65-111.
26. Subramanian SV: **The relevance of multilevel statistical methods for identifying causal neighborhood effects – Commentary.** *Social Science & Medicine* 2004, **58(10):**1961-1967.
27. Osypuk TL, Galea S: **What level macro? Choosing appropriate level to assess how place influences population health.** In *Macrosocial determinants of population health* Edited by: Galea S. New York: Springer; 2007:399-438.
28. Cummins S, Macintyre S, Davidson S, Ellaway A: **Measuring neighbourhood social and material context: generation and interpretation of ecological data from routine and non-routine sources.** *Health & Place* 2005, **11(3):**249-260.
29. Gauvin L, Richard L, Craig CL, Spivock M, Riva M, Forster M, Laforest S, Laberge S, Fournel MC, Gagnon H, *et al.*: **From walkability to active living potential – An "ecometric" validation study.** *American Journal of Preventive Medicine* 2005, **28(2):**126-133.
30. Chaix B, Rosvall M, Lynch J, Merlo J: **Disentangling contextual effects on causespecific mortality in a longitudinal 23-year follow-up study: impact of population density or socioeconomic environment?** *International Journal of Epidemiology* 2006, **35(3):**633-643.
31. Galea S, Ahern J: **Invited commentary: Considerations about specificity of associations, causal pathways, and heterogeneity in multilevel thinking.** *American Journal of Epidemiology* 2006, **163(12):**1079-1082.
32. Openshaw S: **The modifiable areal unit problem.** *Concepts and Techniques in Modern Geography* 1984, **38:**1-41.
33. Openshaw S, Taylor PJ: **A Million or So Correlated Coefficients: Three experiments on the Modifiable Areal Unit Problem.** In *Statistical applications in the spatial sciences* Edited by: Wrigley N, Bennet R. London: Pion; 1979:127-144.
34. Cliff A, Ord J: **Spatial autocorrelation.** In *Monograph No 5 in spatial and environmental systems analysis* Edited by: Chorley R, Harvey D. London: Pion Ltd; 1973:178.
35. Cummins S, Curtis S, Diez-Roux AV, Macintyre S: **Understanding and representing 'place' in health research: A relational approach.** *Social Science & Medicine* 2007, **65(9):**1825-1838.
36. Rothman K, Greenland S: **Modern epidemiology.** 2nd edition. Philadelphia: Lippincott Williams & Wilkins; 1998.
37. Haynes R, Daras K, Reading R, Jones A: **Modifiable neighbourhood units, zone design and residents ' perceptions.** *Health & Place* 2007, **13(4):**812-825.
38. Haynes R, Gale S: **Mortality, long-term illness and deprivation in rural and metropolitan wards of England and Wales.** *Health & Place* 1999, **5(4):**301-312.
39. Blakeley T, Subramanian SV: **Multilevel studies.** In *Methods in social epidemiology* Edited by: Oakes JM, Kaufman JS. San Francisco, CA: Jossey-Bass; 2006:316-340.
40. Statistics Canada Census Operations Division: **2001 Census Dictionary.** Ottawa, Canada: Minister of Industry; 2003.
41. Gauvin L, Riva M, Barnett TA, Richard L, Craig CL, Spivock M, Laforest S, Laberge S, Fournel MC, Gagnon H, *et al.*: **Association between neighborhood active living potential and walking.** *American Journal of Epidemiology* 2008, **167:**944-953.
42. Cervero R: **Mixed land-uses and commuting: Evidence from the American housing survey.** *Transportation Research Part a-Policy and Practice* 1996, **30(5):**361-377.
43. De Bourdeaudhuij I, Sallis JF, Saelens BE: **Environmental correlates of physical activity in a sample of Belgian adults.** *American Journal of Health Promotion* 2003, **18(1):**83-92.
44. Ewing R, Schmid T, Killingsworth R, Zlot A, Raudenbush S: **Relationship between urban sprawl and physical activity, obesity, and morbidity.** *American Journal of Health Promotion* 2003, **18(1):**47-57.
45. Fisher K, Li F, Michael Y: **Neighborhood-level influences on physical activity among older adults: a multi-level analysis.** *Journal of Aging and Physical Activity* 2004, **12(1):**45-63.
46. Frank LD, Pivo G: **Impacts of mixed use and density on utilization of three modes of travel: Single-occupant vehicle, transit, and walking.** *Transportation Research Record* 1995, **1466:**44-52.
47. Frank LD, Schmid TL, Sallis JF, Chapman J, Saelens BE: **Linking objectively measured physical activity with objectively measured urban form – Findings from SMARTRAQ.** *American Journal of Preventive Medicine* 2005, **28(2):**117-125.
48. Giles-Corti B, Broomhall MH, Knuiman M, *et al.*: **Increasing walking: how important is distance to, attractiveness, and size of public open space?** *American Journal of Preventive Medicine* 2005, **28(Suppl 2):**169-176.
49. Giles-Corti B, Donovan RJ: **Socioeconomic status differences in recreational physical activity levels and real and perceived access to a supportive physical environment.** *Preventive Medicine* 2002, **35(6):**601-611.
50. Giles-Corti B, Donovan RJ: **Relative influences of individual, social environmental, and physical environmental correlates of walking.** *American Journal of Public Health* 2003, **93(9):**1583-1589.
51. Lee C, Vernez Moudon A: **Correlates of Walking for Transportation or Recreation Purposes.** *Journal of Physical Activity and health* 2006, **3(Suppl 1):**S77.
52. Pikora TJ, Giles-Corti B, Knuiman MW, Bull FC, Jamrozik K, Donovan RJ: **Neighborhood environmental factors correlated with walking near home: Using SPACES.** *Medicine and Science in Sports and Exercise* 2006, **38(4):**708-714.
53. Saelens BE, Sallis JF, Frank LD: **Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures.** *Annals of Behavioral Medicine* 2003, **25(2):**80-91.
54. Eyler AA, Brownson RC, Bacak SJ, Housemann RA: **The epidemiology of walking for physical activity in the United States.** *Medicine and Science in Sports and Exercise* 2003, **35(9):**1529-1536.
55. Rafferty AP, Reeves MJ, McGee HB, Pivarnik JM: **Physical activity patterns among walkers and compliance with public health recommendations.** *Medicine and Science in Sports and Exercise* 2002, **34(8):**1255-1261.
56. United States Department of Health and Human Services: **Physical Activity and health: A report from the Surgeon General.** Atlanta, GA: U.S: Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion; 1996.
57. Browning CR, Cagney KA: **Moving beyond poverty: Neighborhood structure, social processes, and health.** *Journal of Health and Social Behavior* 2003, **44(4):**552-571.
58. Lochner KA, Kawachi I, Brennan RT, Buka SL: **Social capital and neighborhood mortality rates in Chicago.** *Social Science & Medicine* 2003, **56(8):**1797-1805.
59. Reading R, Langford IH, Haynes R, Lovett A: **Accidents to preschool children: comparing family and neighbourhood risk factors.** *Social Science & Medicine* 1999, **48(3):**321-330.
60. Law M, Wilson K, Eyles J, Elliott S, Jerrett M, Moffat T, Luginaah I: **Meeting health need, accessing health care: the role of neighbourhood.** *Health & Place* 2005, **11(4):**367-377.
61. Lebel A, Pampalon R, Villeneuve PY: **A multi-perspective approach for defining neighbourhood units in the context of a study on health inequalities in the Quebec City region.** *International Journal of Health Geographics* 2007, **6:**27.
62. Martin D, Nolan A, Tranmer M: **The application of zone-design methodology in the 2001 UK Census.** *Environment and Planning A* 2001, **33(11):**1949-1962.
63. Nakaya T: **An information statistical approach to the modifiable areal unit problem in incidence rate maps.** *Environment and Planning A* 2000, **32(1):**91-109.
64. Openshaw S: **Developing GIS-relevant zone-based spatial analysis methods.** In *Spatial analysis: Modelling in a GIS environment* Edited by: Longley P, Batty M. Cambridge: Pearson Professional Ltd; 1996:55-73.
65. Openshaw S, Rao L: **Algorithms for reengineering 1991 census geography.** *Environment and Planning A* 1995, **27(3):**425-446.
66. **Ville de Montréal** [http://www.ville.montreal.qc.ca]
67. ESRI: **ArcGIS 9.2.** Redlands, CA; 2006.
68. Statistics Canada: **2001 Census preview of products and Services.** Ottawa, Canada: Statistics Canada.
69. Theil H: **Statistical decomposition analysis.** Amsterdam: North-Holland; 1972.
70. Theil H, Finezza A: **A note on the measurement of racial integration of schools by means of informational concepts.** *Journal of Mathematical Sociology* 1971, **1:**187-194.
71. Communauté urbaine de Montréal: **Carte d'occupation du sol (édition 2000); document d'accompagnement.** Montréal:

Communauté urbaine de Montréal, division de l'aménagement, service de la mise en valeur du territoire; 2001.

72. Cloutier MS, Apparicio P, Thouez JP: **GIS-based spatial analysis of child pedestrian accidents near primary schools in Montréal, Canada.** *Applied GIS* 2007, **3(4):**.

73. Hewko J, Smoyer-Tomic KE, Hodgson MJ: **Measuring neighbourhood spatial accessibility to urban amenities: Does aggregation error matter?** *Environment and Planning A* 2002, **34(7):**1185-1206.

74. Witten K, Exeter D, Field A: **The quality of urban environments: Mapping variation in access to community resources.** *Urban Studies* 2003, **40(1):**161-177.

75. Apparicio P, Séguin AM: **Measuring the accessibility of services and facilities for residents of public housing in Montréal.** *Urban Studies* 2006, **43(1):**187-211.

76. 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.** *International Journal of Health Geographics* 2008, **7:**7.

77. DMTI Spatial: **CanMap® Streetfiles.** Markham: DMTI Spatial inc; 2005.

78. Tabachnick B, Fidel L: **Using multivariate statistics.** 4th edition. Needham Heights, MA: Allyn & Bacon; 2001.

79. SAS Institute Inc: **SAS version 9.1.** Cary, NC, USA.

80. Harris R, Sleight P, Webber R: **Geodemographics, GIS and neighbourhood targeting.** Chichester, West Sussex, England: John Wiley & Sons, Ltd; 2005.

81. Duda R, Hart P, Stork D: **Pattern classification.** 2nd edition. New York: John Wiley & Sons, Ltd; 2001.

82. Everitt B, Landau S, Leese M: **Cluster analysis.** 4th edition. London: Arnold; 2001.

83. Lebart L, Morineau A, Piron M: **Statistique exploratoire multidimensionnelle.** 2nd edition. Paris: Dunod; 1997.

84. Raudenbush S, Bryk A, Congdon R: **HLM: Hierarchical Linear and Nonlinear Modeling (Version 6.04).** Chicago: Scientific Software International; 2005.

85. Lee J, Wong DWS: **Statistical analysis with ArcView GIS.** New York: John Wiley & Sons, Inc; 2001.

86. Longley P, Batty M: **Spatial analysis: Modelling in a GIS environment.** Cambridge: earson Professional Ltd; 1996.

87. Milligan GW, Cooper MC: **An examination of procedures for determining the number of clusters in a data set.** *Psychometrika* 1985, **50(2):**159-179.

88. Calinski T, Harabasz J: **A dendrite method for cluster analysis.** *Communications in Statistics* 1974, **3(1):**1-27.

89. SAS Institute Inc: **SAS Technical Report A-108, Cubic clustering criterion.** Cary (NC), USA: SAS Institute Inc:54.